IDS Reference L.

[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 17/30
G10H 1/00 G10L 15/02
G10L 15/20

[12] 发明专利申请公开说明书

[21] 申请号 01813565.X

[43] 公开日 2005 年 3 月 9 日

[11] 公开号 CN 1592906A

[22] 申请日 2001.7.26 [21] 申请号 01813565.X
[30] 优先权
[32] 2000.7.31 [33] US [31] 60/222,023
[32] 2001.4.20 [33] US [31] 09/839,476
[86] 国际申请 PCT/EP2001/008709 2001.7.26
[87] 国际公布 WO2002/011123 英 2002.2.7
[85] 进入国家阶段日期 2003.1.29
[71] 申请人 沙扎姆娱乐有限公司
地址 英国伦敦
[72] 发明人 埃弗里·L·C·王
朱利叶斯·O·史密斯第三

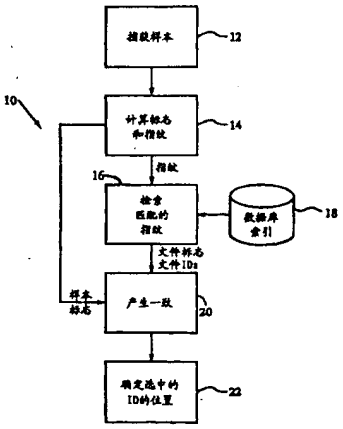
[74] 专利代理机构 北京市柳沈律师事务所
代理人 邸万奎 黄小临

权利要求书 8 页 说明书 25 页 附图 13 页

[54] 发明名称 用于在强噪声和失真下识别声音和音乐信号的系统和方法

[57] 摘要

识别音频样本方法，从索引大组原始记录的数据库中，定位音频样本匹配的音频文件。 数据库索引中，每索引的音频文件由标志时间点和关联指纹代表。 标志出现在文件内可再生位置，指纹代表标志时间点处、附近信号特征。 为执行识别，未知的样本计算标志和指纹，使用标志和指纹从数据库检索匹配指纹。 对每包含匹配指纹的文件，标志与样本的计算出相同指纹的标志相比较。 若大量对应标志线性相关，样本和检索文件等价指纹时间演化相同，文件认为与样本同。 此法可用于任何声音、音乐，对遭受背景噪声、压缩人工信号、传送信息遗失的线性或非线性失真音频信号特有效。 样本可在与数据库项目数对数成比例时间内辨认；给定充分计算力，可随声音采样实时执行识别。



ISSN1008-4274

知识产权出版社出版

1. 一种用于比较媒体样本和媒体文件的方法，包括:

确定一组样本指纹，每个样本指纹描述所述的媒体样本内的一个特定的位置;

获取一组文件指纹，每个文件指纹描述所述的媒体文件中至少一个文件位置;

产生所述的媒体样本的所述的特定的位置与所述的文件的所述的文件位置之间的一致（correspondence），其中，相对应的位置具有等价的指纹;

如果多个所述的相对应的位置充分线性相关，则辨认所述的媒体样本和所述的媒体文件.

2. 一种用于比较音频样本和音频文件的方法，包括:

对至少一个音频文件中的每一个，计算多个代表所述的音频文件的指纹;

计算多个代表所述的音频样本的样本指纹; 以及

如果至少一个阈值数目的所述的文件指纹等价于所述的样本指纹，则辨认所述的音频样本和所述的媒体文件;

其中，所述的样本指纹对所述的音频样本的时间展宽不变。

3. 一种用于描述音频样本特征的方法，包括:

在所述的音频样本中计算一组可再生的位置; 以及

在所述的音频样本中计算一组描述所述的可再生位置的指纹。

4. 一种描述音频样本特征的方法，包括从所述的音频样本的声频谱图计算之少一个指纹，其中，所述的声频谱图包括锚凸（salient）点和链接的凸点，且其中，所述的指纹是从所述的锚凸点和任何链接的凸点的频率坐标计算出的。

5. 一种用于识别媒体样本的方法，包括辨认这样的媒体文件，其中，所述的媒体文件与所述的媒体样本的充分多个的等价特征的位置充分地线性相关。

6. 一种用于识别媒体文件的方法，包括:

对于多个媒体文件中的每一个，提供所述的媒体文件的文件代表;

提供所述的媒体样本的样本代表; 以及

通过搜索所述的文件代表，在所述的文件代表中辨认至少一个相似的文

件代表，其中，所述的相似的文件代表相似于所述的样本代表，其中，所述的搜索的执行部分地依赖于所述的文件代表的辨认的概率，

7. 一种用于识别媒体样本的方法，包括：

计算一组描述一段所述的媒体样本的样本指纹；

在滚动（rolling）缓冲中存储所述的指纹；

在数据库索引中获取一组匹配的指纹，每个匹配的指纹描述至少一个媒体文件并与所述的滚动缓冲中的至少一个指纹相匹配；

辨认至少一个具有多个匹配的指纹的媒体文件；以及

从所述的滚动缓冲中移除至少一个样本指纹。

8. 一种计算机实现的方法，用于创建数据库中至少一个音频文件的数据库索引，包括：

计算一组代表每个音频文件的特征的指纹，每个指纹描述所述的音频文件内特定的位置；以及

在存储器内存储所述的指纹、所述的位置、和每个媒体文件的标识符，其中，在所述的存储器中，每个相对应的指纹、位置与标识符是相关联的。

9. 如权利要求 1 的方法，其中，所述的确定步骤包括计算一组样本指纹。

10. 如权利要求 1 的方法，其中，所述的确定步骤包括接收一组样本指纹。

11. 一种用于比较音频样本和音频文件的方法，包括：

对于至少一个音频文件中的每一个，计算多个代表所述的音频文件的文件指纹；

计算多个代表所述的音频样本的样本指纹；以及

如果至少一个阈值数目的所述的文件指纹等价于所述的样本指纹，则辨认所述的媒体样本和所述的媒体；

其中，根据权利要求 4 的方法，每个样本指纹从所述的音频样本的声频谱图计算出。

12. 如权利要求 9 的方法，其中，所述的媒体样本是音频样本。

13. 如权利要求 9 的方法，其中，所述的辨认步骤包括在所述的相对应的位置的散布图中确定对角线的位置。

14. 如权利要求 13 的方法，其中，确定所述的对角线的位置包括求所述的相对应的位置之间的差值。

15. 如权利要求 14 的方法，其中，确定所述的对角线的位置还包括对所述的差值分类。

16. 如权利要求 14 的方法，其中，确定所述的对角线的位置还包括计算所述的差值的柱状图的尖峰（peak）。

17. 如权利要求 9 的方法，其中，所述的辨认步骤包括计算所述的一致的 Hough 变换或 Radon 变换。

18. 如权利要求 17 的方法，其中，所述的辨认步骤还包括确定所述的 Hough 或 Radon 变换的尖峰的位置。

19. 如权利要求 9 的方法，其中，所述的辨认步骤包括确定所述的一致的数目是否超过了阈值数目。

20. 如权利要求 9 的方法，还包括:

从数据库索引中获取描述附加的媒体文件的文件位置的附加的指纹;

产生所述的媒体样本的所述的特定的位置与所述的附加的媒体文件的所述的文件位置之间的附加的一致，其中，相对应的位置具有等价的指纹；以及

选择选中的媒体文件，其中，所述的选中的媒体文件具有最多个的充分线性相关的一致位置。

21. 如权利要求 20 的方法，还包括辨认有多个所述的相对应的位置充分线性相关的媒体文件，且其中，所述的选择步骤包括从所述的被辨认的媒体文件中选择选中的媒体文件。

22. 如权利要求 21 的方法，其中，辨认有多个所述的相对应的位置充分线性相关的所述的媒体文件包括搜索所述的附加的媒体文件的第一子集。

23. 如权利要求 22 的方法，其中，所述的第一子集中的附加的媒体文件比不在所述的第一子集中的附加的媒体文件有较高的被辨认的概率。

24. 如权利要求 22 的方法，其中，辨认有多个所述的相对应的位置充分线性相关的所述的媒体文件还包括搜索所述的附加的媒体文件的第二子集，其中，如果在所述的第一子集中没有确认出媒体文件，则搜索所述的第二子集。

25. 如权利要求 21 的方法，还包括根据被辨认的概率对所述的附加的媒体文件排列的顺序。

26. 如权利要求 25 的方法，其中，辨认有多个所述的相对应的位置充分

线性相关的所述的媒体文件包括根据所述的排序搜索所述的附加的媒体文件。

27. 如权利要求 21 的方法，其中，辨认有多个所述的相对应的位置充分线性相关的所述的媒体文件包括：在具有超过预定的阈值数目的多个所述的充分线性相关的相对应的位置的媒体文件处终止搜索。

28. 如权利要求 9 的方法，其中，所述的方法在分布式的系统中实现。

29. 如权利要求 28 的方法，其中，所述的计算步骤在客户设备中执行，所述的获取、产生、以及辨认步骤在中心位置执行，且本方法还包括从所述的客户设备向所述的中心位置传送所述的样本指纹。

30. 如权利要求 9 的方法，还包括对所述的媒体样本的顺序增长段重复所述的计算、获取、产生、以及辨认步骤。

31. 如权利要求 9 的方法，其中，所述的获取、产生、以及辨认步骤以周期性间隔对存储所述的计算出的指纹的滚动缓冲执行。

32. 如权利要求 9 的方法，还包括获取所述的媒体文件，其中，所述的计算步骤和所述的获取步骤同时执行。

33. 如权利要求 8 的方法，还包括按指纹值对所述的数据库索引分类。

34. 如权利要求 8 或 9 的方法，其中，每个音频文件或所述的媒体样本内的所述的特定的位置依靠所述的音频文件或媒体样本计算出。

35. 如权利要求 8 或 9 的方法，其中，每个指纹代表在所述的特定的位置附近的所述的音频文件或媒体样本的至少一个特征。

36. 如权利要求 8 或 9 的方法，其中，所述的指纹是数字值。

37. 如权利要求 8 或 12 的方法，其中，所述的指纹的值指定了用于计算所述的指纹的方法。

38. 如权利要求 8 或 12 的方法，其中，所述的特定的位置是所述的音频文件或样本中的时间点。

39. 如权利要求 38 的方法，其中，所述的时间点出现在所述的音频文件或样本的声频谱 Lp 范数（norm）的局部最大值处。

40. 如权利要求 8 或 12 的方法，其中，所述的指纹从对所述的音频文件或样本的频率分析计算出。

41. 如权利要求 8 或 12 的方法，其中，所述的指纹是从由频谱段（slice）指纹、线性预测编码系数、和倒频谱(cepstral)系数组成的组中选择出的。

42. 如权利要求 8 或 12 的方法，其中，所述的指纹从所述的音频文件或样本的声频谱图计算出。

43. 如权利要求 42 的方法，其中，所述的声频谱图的凸点包括时间坐标和频率坐标，且其中，所述的特定的位置从所述的时间坐标计算出，而所述的指纹从所述的频率坐标计算出。

44. 如权利要求 43 的方法，还包括将多个所述的凸点连接到锚凸点，其中，一个所述的特定的位置从所述的锚凸点的时间坐标计算出，而相对应的指纹从至少一个所述的链接的凸点和所述的锚点的频率坐标计算出。

45. 如权利要求 44 的方法，其中，所述的相对应的指纹从所述的链接的凸点和所述的锚点的两个所述的频率坐标之商计算出，从而所述的相对应的指纹为时间展宽不变。

46. 如权利要求 45 的方法，其中，所述的相对应的指纹还从所述的锚点的所述的时间坐标和所述的链接的凸点的时间坐标之间的至少一个时间差值计算出。

47. 如权利要求 46 的方法，其中，所述的相对应的指纹还从一个所述的时间差值和所述的链接的凸点和所述的锚点的一个所述的频率坐标的乘积计算出，从而所述的相对应的指纹为时间展宽不变。

48. 如权利要求 8 或 12 的方法，其中，所述的特定的位置与所述的指纹从所述的音频文件或样本的多维函数的凸点计算出，其中，至少一个所述的维是时间维，且至少一个所述的维是非时间维。

49. 如权利要求 48 的方法，其中，所述的特定的位置从所述的时间维计算出。

50. 如权利要求 48 的方法，其中，所述的指纹从至少一个所述的非时间维计算出。

51. 如权利要求 11 或 48 的方法，其中，所述的凸点从由所述的多维函数的局部最多、局部最少、和零交叉组成的组中选择。

52. 如权利要求 8 或 12 的方法，其中，所述的指纹为时间展宽不变。

53. 如权利要求 8 或 12 的方法，其中，每个指纹从所述的音频文件或样本的多个时间段计算出。

54. 如权利要求 53 的方法，其中，所述的多个时间段被偏移可变的时间量。

55. 如权利要求 54 的方法，其中，所述的指纹部分地从所述的可变的量计算出。

56. 如权利要求 6 的方法，其中，所述的至少一个相似的文件代表对所述的样本代表超过阈值相似性。

57. 如权利要求 6 的方法，其中，所述的辨认步骤包括搜索所述的文件代表的第一子集，其中，所述的第一子集包含具有比不在所述的第一子集中的文件代表高的辨认的概率的文件代表。

58. 如权利要求 57 的方法，还包括：如果所述的第一子集不包括所述的至少一个相似的文件代表，则搜索所述的文件代表的第二子集。

59. 如权利要求 6 的方法，还包括按所述的被辨认概率对所述的文件代表排列的顺序，其中，所述的辨认步骤包括按所述的排列的顺序搜索所述的文件代表。

60. 如权利要求 59 的方法，还包括：在辨识到所述的至少一个相似的文件代表时终止所述的搜索。

61. 如权利要求 6、23 或 25 的方法，其中，所述的被辨识概率部分依靠先前辨认的新旧程度（recency）计算出。

62. 如权利要求 61 的方法，其中，在所述的特定的文件代表被辨认时，特定的文件代表的新旧程度分值增加。

63. 如权利要求 61 的方法，其中，所述的文件代表的新旧程度分值以规则的时间隔降低。

64. 如权利要求 63 的方法，其中，所述的新旧程度分值随时间指数地降低。

65. 如权利要求 6 或 23 的方法，其中，辨认的所述的概率部分地依靠先前识别的频率而计算出。

66. 如权利要求 2 的方法，其中，所述的样本指纹包括所述的音频样本的频率分量的商。

67. 如权利要求 2 的方法，其中，所述的样本指纹包括所述的音频样本的频率分量与所述的音频样本中的点之间的时间差值的乘积。

68. 如权利要求 4、11 或 44 的方法，其中，所述的链接的凸点落在目标区域内。

69. 如权利要求 68 的方法，其中，所述的目标区域按时间范围定义。

70. 如权利要求 68 的方法，其中，所述的目标区域按频率范围定义。

71. 如权利要求 68 的方法，其中，所述的目标区域可变。

72. 如权利要求 7 的方法，还包括对所述的媒体样本的附加的段重复所述的方法。

73. 如权利要求 7 的方法，其中，所述的计算、存储、和移除步骤在客户设备中执行，而所述的获取和辨认步骤在中心位置执行，且其中，该方法还包括将所述的样本指纹从所述的客户设备传送到所述的中心位置。

74. 如权利要求 7 的方法，其中，所述的计算步骤在客户设备中执行，而所述的存储、获取、辨认、和移除步骤在中心位置执行，且其中，该方法还包括将所述的样本指纹从所述的客户设备传送到所述的中心位置。

75. 如权利要求 3 的方法，其中，所述的可再生的位置与所述的指纹同时计算。

76. 一种计算机可存取的程序存储设备，切实地包含可由所述的计算机执行的指令的程序，以执行用于比较媒体样本和媒体文件的方法步骤，所述的方法步骤包括:

计算一组样本指纹，每个样本指纹描述所述的媒体样本内的特定的位置;

获取一组文件指纹，每个文件指纹描述所述的媒体文件中的至少一个文件位置;

产生所述的媒体样本的所述的特定的位置与所述的媒体文件的所述的文件位置之间的一致，其中，相对应的位置具有等价的指纹;

如果多个所述的相对应的位置充分地线性相关，则辨认所述的媒体样本和所述的媒体。

77. 一种用于识别媒体样本的系统，包括:

标志处理和指纹理对象，用于计算在所述的媒体样本内的一组特定的位置与一组样本指纹，每个样本指纹描述一个所述的特定的位置;

数据库索引，包含至少一个媒体文件的文件位置与相对应的指纹; 以及

分析对象，用于:

在所述的数据库索引中确定一组匹配的指纹的位置，其中，所述的匹配的指纹等价于所述的样本指纹;

产生所述的媒体样本中的所述的特定的位置与所述的至少一个媒体文件中的文件位置之间的一致，其中，相对应的位置具有等价的指纹; 以及

辨认至少一个媒体文件，其中，多个所述的相对应的位置充分地线性相关。

用于在强噪声和失真下识别声音和音乐信号的系统和方法

技术领域

本发明大体涉及基于内容的信息检索。更具体地说，本发明特别涉及音频信号的识别，所述的音频信号包括高度失真的、或包含强噪声的声音或音乐。

背景技术

越来越需要自动识别从多种来源产生的音乐或其它音频信号。例如，有版权的作品的拥有者或广告人员对于获取关于其材料的广播频率的数据感兴趣。音乐跟踪服务在大市场中提供主要无线电台的节目表。消费者希望辨认广播中的歌曲或广告，以便可以购买新的、有趣的音乐或其它产品和服务。当其由人工执行时，任何种类的持续的或点播(on-demand)的声音识别都是低效且费力的。这样，识别音乐或声音的自动方法将给消费者、艺术家、以及多种产业带来重大的意义。随着音乐发行模式从店铺购买转移到了经因特网下载，将用计算机实现的音乐识别和因特网购买以及其它基于因特网的服务直接连接起来是非常可行的。

传统上，对广播中播放的歌曲的识别，是通过使播放歌曲的无线电台和时间，与无线电台或第三方来源提供的节目表相匹配来执行的。这种方法内在地限定于可获取信息的无线电台。其它方法则依赖于在广播信号中嵌入不可听的码。被嵌入的信号在接收器中解码，以抽取关于广播信号的辨认信息。这种方法的缺点在于需要专用的解码设备以辨认信号，而且只能辨认那些具有嵌入码的歌曲。

任何大规模音频识别都需要某种基于内容的音频检索，其中，未辨认的广播信号与已知信号的数据库比较，以辨认相似的和相同的数据库信号。需要注意，基于内容的音频检索不同于现有的通过网络搜索引擎的音频检索，其中，只搜索围绕音频文件或与音频文件相关联的后数据(metadata)文本。还需要注意，尽管语音识别对于将有声的信号转变成可以使用公知技术来索引和搜索的文本很有用，但是其不适用于包含音乐和声音的大多数音频信号。

在某些方面，音频信息检索类似于由搜索引擎提供的基于文本的信息检索。在其它方面，音频识别并不类似于：音频信号缺乏可简单地辨认的诸如文字之类的实体，所述的实体提供用于搜索或索引的标识符。同样地，当前的音频检索方案通过计算出的代表信号的各种品质和特征的知觉特征进行索引。

典型地，基于内容的音频检索通过分析查询信号来执行，以获取许多代表性的特征，然后对所得特征进行相似性测量以确定最相似于该查询信号的数据库文件的位置。所接收的对象的相似性必然是所选择的知觉特征的反映。本领域有很多基于内容的检索方法可用。例如，发布到 Kenyon 的美国专利第 5,210,820 号公开了一种信号识别方法，其中，所接收信号被处理并采样以获取每个采样点的信号值。然后，计算采样的值的统计动差，以产生可以与所存储信号的标识符比较的特征矢量，来检索相似的信号。发布到 Kenyon 以及其他地方的美国专利第 4,450,531 号和第 4,843,562 号公开了相似的广播信息分类方法，其中，计算了未辨认信号与所存储的参考信号之间的交叉相关性。

在 J. T. Foote，“Content-Based Retrieval of music and Audio (音乐和音频的基于内容的检索)”，在 C.-C. J. Kuo et al.，editor，Multimedia Storage and Archiving Systems II(多媒体存储和存档系统 II)，Proc. of SPIE，volume 3229，pages 138-147，1997 中公开了一种用于通过声学相似性检索音频文件的系统。通过将每个音频文件参数化为以唛耳标度的(mel-scaled)倒频谱(cepstral)系数来计算特征矢量，并且从该参数化数据生成量化树(quantization tree)。为执行查询，未知的信号被参数化，以获取特征矢量，而该特征矢量被分类为树上的叶节点(leaf node)。为每个叶节点收集柱状图，从而产生代表该未知的信号的 N 维矢量。两个这样的矢量之间的距离表示两个声音文件之间的相似性。在这种方法中，基于人们在其中分配了训练数据(training data)的类(class)，被监管的量化方案知道区分音频特征，而忽略不重要的变化。依靠分类系统，选择不同的声学特征为重要特征。这样，不仅来识别音乐，这种方法更适合于发现歌曲之间的相似性并将音乐分类。

发布到 Blum 以及其他地方的美国专利第 5,918,223 公开了一种基于内容对音频信息的分析、存储、检索、及分段方法。在这种方法中，在每个文件的周期性间隔测量许多诸如音量(loudness)、低音(bass)、音调(pitch)、亮度(brightness)、带宽(bandwidth)、唛耳-频率(mel-frequency) 倒频谱系数之类的声学特征。将这些特征进行统计测量并结合以形成特征矢量。基于数据库中

的音频数据文件的特征矢量到未辨认的文件的特征矢量的相似性，检索数据库中的音频数据文件。

所有上述现有的技术的音频识别方法的关键问题在于，当要识别的信号遭受由于，例如，背景噪声、传输错误和信息遗失(dropout)、干扰、带宽受限制的滤波(band-limited filtering)、量化、时间变形(time-warping)、以及语音质量数字压缩而造成的线性和非线性失真时，就容易失效。在现有的技术方法中，当处理失真的声音样本以获取声学特征时，只能发现一部分从原始记录导出的特征。因此，结果特征矢量与原始记录的特征矢量并不非常相似，未必能执行正确的识别。仍然需要一种声音识别系统，能在强噪声和失真的条件下良好地工作。

现有的技术方法的另一个问题在于，其计算量大，且不能良好地分等级(scale)。这样，不可能使用现有的技术方法用大型数据库进行实时识别。在这种系统中，不可以使数据库具有多于几百或上千条的记录。现有的技术方法中的搜索时间倾向于随着数据库的大小线性增长，这使得对上百万的声音记录进行分等级在经济上是不可行的。Kenyon 的方法同样需要大量的专用的数字信号处理硬件。

现有的商用方法通常对能够执行识别的输入样本有严格的要求。例如，其要求完整的歌曲，或至少 30 秒的歌曲，以便采样，或者要求歌曲从头采样。其还难于识别在单个流(stream)中混合在一起的多个歌曲。所有这些缺点使得现有的技术的方法在许多实际应用中的使用是不可行的。

发明内容

相应地，本发明的主要目的是提供一种方法，用于识别遭受强噪声和失真的音频信号。

本发明的再一个目的是提供一种识别方法，其可以只基于将要被辨认的信号的几秒钟而实时地执行。

本发明的另一个目的是提供一种识别方法，其可以基于声音中几乎任何位置的样本识别声音，而不仅仅是只能在开头。

本发明的一个附加的目的是提供一种识别方法，其不需要使声音样本被编码或与特定的无线电台或节目表相关联

本发明的再一个目的是提供一种识别方法，其可以识别在单个流中混合

在一起的多个歌曲中的每一首歌曲。

本发明的另一个目的是提供一种声音识别系统，在其中，可以通过事实上任何已知方法从任何环境向所述的系统提供未知的声音。

这些目的和优点可以通过给定许多已知媒体文件的数据库索引，用一种用于识别诸如音频样本之类的媒体样本的方法来获得。数据库索引包含代表被索引的媒体文件的特定的位置处的特征的指纹(fingerprint)。未知的媒体样本与数据库中的一个媒体文件(选中的(winning)媒体文件)是一致的，该媒体文件指纹相对位置与样本的指纹的相对位置最为匹配的。在音频文件的情况中，选中的文件的指纹的时间演化（evolution）与样本中的指纹的时间演化相匹配。

本方法最好在分布的计算机系统中实现，并包含如下步骤：在样本的特定的位置确定一组指纹；在数据库索引中确定匹配的指纹的位置；产生在样本中的位置与具有等价的指纹的文件中的位置之间的一致；以及辨认非常多个一致充分线性相关的媒体文件。具有最大数目的线性相关的一致的文件被认为是选中的媒体文件。辨认具有大量的一致的文件的一种方法是：执行等价于扫描从多对一致中产生的散布(scatter)图中的对角线的过程。在一个实施例中，辨认具有大量的线性一致的媒体文件包括只搜索媒体文件的第一子集。第一子集中的文件比不在第一子集中的文件有较高的被辨认的概率。辨认的概率最好基于实验频率或先前辨认的新旧程度测量，连同对辨认频率的推理(a priori)预测。如果第一子集中没有媒体文件被辨认，则搜索包含其余文件的第二子集。作为选择，文件可以按概率排列，并以排列的顺序搜索。当文件的位置被确定时终止搜索。

最好是，样本内的特定的位置以独立于样本的方式可再生地计算出。这样的可再生可计算位置被称为“标志(landmark)”。指纹最好是数值。在一个实施例中，每个指纹代表在每个位置或从该位置微小的偏移处的媒体样本的多个特征。

本方法对识别音频样本特别有用，其中，特定的位置是音频样本中的时间点。这些时间点出现在，例如，音频样本的频谱 Lp 范数(norm)的局部最大值处。指纹可以通过对音频样本的任何分析计算出，并且最好相对于样本的时间展宽(time stretching)不变。指纹的例子包括频谱段指纹、多段指纹、线性预测编码 (LPC) 系数、倒频谱(cepstral)系数、以及声频谱图(spectorgram)尖

峰的频率分量。

本发明还提供一种用于实现上述方法的系统，包含：标志处理(landmarking)对象，用于计算特定的位置；指纹处理(fingerprinting)对象，用于计算指纹；数据库索引，包含针对媒体文件的文件位置与指纹；以及一个分析对象。分析对象，通过在数据库索引中确定匹配的指纹的位置，产生一致，并分析一致，以选择选中的媒体文件，来实现本方法。

还提供一种可由计算机存取的程序存储设备，切实地包含可由计算机执行的指令的程序，以执行针对上述方法的方法步骤。

另外，本发明提供一种用于在数据库中创建许多音频文件的索引的方法，包含下述步骤：在每个文件的特定的位置计算一组指纹；并存储指纹、位置、以及存储器中的文件的标识符(identifier)。在存储器中，相对应的指纹、位置、标识符被关联起来形成一个三元组(triplet)。最好是，可以是音频文件内的时间点的位置，依赖于文件而被计算，并可再生。例如，时间点可以出现在音频文件的频谱 Lp 范数的局部最大值处。在某些情况中，最好是数值的每个指纹，代表特定的位置附近的文件的许多特征。可以从任何对音频文件的分析或数字信号处理来计算出指纹。指纹的例子包括频谱段指纹、多段指纹、线性预测编码系数、倒频谱(cepstral)系数、声频谱图尖峰的频率分量、以及被连接的声频谱图尖峰。

最后，本发明提供：用于辨认并入了相对于时间展宽不变的指纹的音频样本的方法，以及用于各种分级搜索的方法

附图说明

图 1 是用于识别声音样本的本发明的方法的流程图。

图 2 是用于实现图 1 的方法的示范性的分布式的计算机的方框图。

图 3 是用于建立在图 1 的方法中使用的声音文件的数据库索引的方法的流程图。

图 4 概略地说明了为声音样本计算的标志和指纹。

图 5 是针对声音样本的 L4 范数的图，说明了标志的选择。

图 6 是用于建立在图 1 的方法中使用的声音文件的数据库索引的可选的实施例的流程图。

图 7A-7C 示出了具有标明的凸点(salient point)和链接的凸点的声频谱图。

图 8A-8C 说明了图 3 的方法的索引集合、索引列表、和主索引列表。

图 9A-9C 说明了图 1 的方法的索引列表、候选列表、和散布列表(scatter list)。

图 10A-10B 分别是说明未知的声音样本的正确辨认和辨认不足的散布图。

具体实施方式

本发明提供一种用于在已给定包含大量的已知媒体文件的数据库的情况下，识别外来的媒体样本的方法。还提供一种用于产生数据库索引的方法，所述的数据库索引允许使用本发明的识别方法进行有效搜索。尽管下述讨论主要涉及音频数据，应该理解，本发明的方法可以适用于任何类型的媒体样本和媒体文件，包括但不限于：文本、音频、视频、图像、和单个媒体类型的任何多媒体组合。在音频的情况中，本发明对识别包含高度线性或非线性失真的样本特别有用，其中，所述的失真是由于，例如，背景噪声、传输错误和信息遗失、干扰、带宽受限制的滤波、量化、时间变形、以及语音质量数字压缩所造成的。随着从下文的描述中将变得清晰，本发明之所以在这样的条件下起作用，是因为：即使只有小一部分计算出的特征幸免于失真，其也可以正确地识别失真的信号。通过本发明可以识别任何类型的音频信号，包括声音、语音、音乐、或多个类型的结合。音频样本的例子包括所记录的音乐、无线广播节目、以及广告。

如这里所使用的，外来的媒体样本是从如下文中所描述的多种来源获取的任意大小的媒体数据的片断。为了执行识别，样本必须是已在本发明所使用的数据库中索引的媒体文件的一部分的再现。所述的被索引的媒体文件可以看作是原始记录，而样本则为原始记录的失真和/或删节的版本或者再现。典型地，样本只与被索引文件的一小部分一致。例如，可以对在数据库中索引的五分钟长的歌曲的十秒钟片断执行识别。尽管用术语“文件”描述被索引的实体，但是所述的实体可以是任何能够获取必要值(如下所述的)的形式。而且，在获取该值后，不需要存储或访问该文件。

图 1 示出了概念上说明本发明的方法 10 的全部步骤的方框图。下文中更详细地描述了各个步骤。本方法辨认选中的媒体文件，即一种其特征指纹的相对位置与外来的样本的同样的指纹的相对位置最接近地匹配。在步骤 12 中

捕获到外来的样本之后，就在步骤 14 中计算标志和指纹。标志出现于样本中特定的位置，即，时间点。标志在样本中的位置最好由样本自身确定，即，依赖于样本品质，并且是可再生的。也就是说，每次重复处理时，为相同的信号计算相同的标志。对于每一个标志，指纹在所获标志处或其附近描述样本的一个或多个特征。特征与标志的接近程度(nearness)通过所使用的指纹处理方法来定义。在某些情况中，如果特征明显地与一个标志一致而与先前或随后的标志不一致，则认为该特征接近于该标志。在其它情况中，特征与多个邻近的标志一致。例如，文本指纹可以是字串(word string)；音频指纹可以是声频谱(spectral)分量；而图像指纹可以是像素红绿蓝(RGB)值。下文中描述了步骤 14 的两个一般实施例，一个实施例中依次计算标志和指纹，而另一个实施例中同时计算标志和指纹。

在步骤 16 中，样本指纹被用来检索存储在数据库索引 18 中的多组匹配的指纹，在所述的数据库索引 18 中，匹配的指纹与一组媒体文件的标志和标识符相关联。然后，使用该组被检索的文件标识符和标志值，来产生一致对(correspondence pair)(步骤 20)，所述的一致对包含样本标志(在步骤 14 中计算出的)和被检索的文件标志，在此计算出了相同的指纹。然后，作为结果的一致对按歌曲标识符分类，为每个可用的文件产生样本标志与文件标志之间的多组一致。扫描每一组，以进行文件标志与样本标志之间的校准。也就是说，辨认各对标志中的线性一致，并根据线性相关的对的数目对该组评分。当大量相对应的样本位置与文件位置可以在一定的容限内用充分相同的线性方程描述时，就出现线性一致。例如，如果描述一组一致对的多个方程的斜率在±5%范围内变化，那么该整组一致可以看作是线性相关的。当然，可以选择任何合适的容限。具有最高分，即具有最大量的线性相关的一致的组的标识符，是选中的文件标识符，被确定其位置，并在步骤 22 中返回。

如下文中进一步描述的，可以用与数据库中实体的数目的对数成正比的时间分量执行识别。基本上，可以实时地执行识别，即使用很大的数据库。也就是说，在获取样本之时，以小的时间滞后，就可以识别样本。本方法可以基于 5-10 秒，甚至低至 1-3 秒的片断辨认声音。在一个优选实施例中，随着在步骤 12 中捕获样本，实时执行标志处理和指纹处理分析，即步骤 14。当样本指纹变为可用时，就执行数据库查询(步骤 16)，并积累一致结果，周期性地扫描线性一致。这样，本方法的所有步骤同时发生，而非图 1 中所建

议的依次线性样式。需要注意的是，本方法与文本搜索引擎是部分地相似的：用户提交查询样本，并返回在声音数据库中索引的匹配文件。

典型地，本方法作为在计算机上运行的软件实现，其中，各个步骤作为独立的软件模块最有效地实现。这样，实现本发明的系统可以认为由标志处理和指纹处理的对象、被索引的数据库、和分析对象组成，用于搜索数据库索引，计算一致，并辨认选中的文件。在依次标志处理和指纹处理的情况中，标志处理和指纹处理的对象可以被认为是不同的标志处理和指纹处理的对象。用于不同对象的计算机指令代码存储在一个或多个计算的存储器中，并由一个或多个计算机处理器执行。在一个实施例中，代码对象与诸如基于英特尔的(intel-based)个人计算机或其它工作站之类的单个计算机系统集群(cluster)在一起。在一个优选实施例中，本方法是通过中央处理器(CPU)的网络的集群来实现的，其中，不同的处理器执行不同的软件对象，以便分散计算量。作为选择，每个 CPU 可以有所有软件对象的副本，允许全同配置的元件的对等网络(homogeneous network)。在这种后者的配置中，每一个 CPU 具有数据库索引的子集，并负责搜索其自己的媒体文件的子集。

尽管本发明不限于任何特定的硬件系统，图 2 中概略地说明了分布式的计算机系统 30 的一个优选实施例的例子。系统 30 包含一个集群的基于 Linux 的 (Linux-based) 处理器 32a-32f，这些处理器是通过多处理总线结构 (multiprocessing bus architecture) 34 或诸如 Beowulf 集群计算机协议之类的连网协议，或两者的混合，来连接的。在这样的安排下，数据库索引最好存储在集群中的至少一个节点 32a 上的随机存取存储器(RAM)中，以确保非常迅速地进行指纹搜索。与其它对象相对应的诸如标志处理节点 32c 和 32f、指纹处理节点 32b 和 32e、以及校准扫描节点 32d 之类的计算节点，不需要与支持数据库索引的节点或多个节点 32a 一样多的随机存取存储器。这样，指定给每个对象的计算节点的数目可以根据需要而调节，使得没有单个对象成为瓶颈。所以，计算网络是高度可并行的，且可以额外地处理被分布在可用的计算资源中的多个同时信号识别查询。这表明，这使得大量的用户可以请求识别并接近实时地接收结果的应用成为可能。

在一个作为选择的实施例中，某些功能对象会更紧密地耦合在一起，而与其它对象保持较不紧密的耦合。例如，标志处理和指纹处理对象可以存在于与其它计算对象在物理上分离的位置。一个这种例子是标志处理和指纹处

理对象与信号捕获处理的紧密联合。在这种安排下，标志处理和指纹处理对象可以作为要嵌入的附加的硬件或软件并入，例如，移动电话、无线应用协议(WAP)浏览器、个人数字助理(PDA)、或其它诸如音频搜索引擎的客户端之类的远程终端。在基于因特网的诸如内容标识服务之类的音频搜索服务中，标志处理和指纹处理对象可以并入客户浏览器应用程序中，作为软件指令或诸如微软动态连接库(DLL)之类的软件插入模块的被连接的组。在这些实施例中，所结合的信号捕获、标志处理、以及指纹处理对象，构成了该服务的客户端。客户端向服务器端发送所捕获的信号样本的抽取特征(feature-extracted)的摘要，所述的信号样本包含标志和指纹对，而服务器端执行该识别。向服务器端发送这种抽取特征的摘要而不是未加工的捕获的信号是有利的，因为大大地减少了数据量，通常以 500 或更大的因数减少。这样的信息，可以通过低带宽侧信道，连同或代替例如发送到服务器的音频流，被实时地发送。这使得能够在公共通讯网络上执行本发明，所述的公共通讯网络为每个用户提供相对小的带宽。

现在将参考音频样本和在声音数据库中索引的音频文件来描述本方法。本方法由两个主要的成分构成，即声音数据库索引构建和样本识别。

数据库索引构建

在可以执行声音识别之前，必须构建可搜索的声音数据库索引。如这里所使用的，数据库是数据的任意索引的集合，而且不限于商业可用的数据库。在数据库索引中，数据的相关元素彼此关联，且每个元素可以被用于检索所关联的数据。声音数据库索引包含：针对记录的所选择的集合或库中的每个文件或记录的索引集合，所述的记录包括演讲、音乐、广告、声纳签名(sonar signature)、或其它声音。每个记录也具有唯一的标识符、声音_ID (sound_ID)。声音数据库本身不需要为每个记录存储音频文件，但是声音_ID 可以被用于检索来自别处的音频文件。期望声音数据库索引非常大，包含针对数百万或甚至上亿的文件的索引。新记录最好以递增的方式添加到数据库索引中。

图 3 中示出了用于根据第一个实施例来构建可搜索声音数据库索引的优选方法 40 的方框图。在本实施例中，首先计算标志，然后在标志处或其附近计算指纹。本领域中一般技术人员将会明白，可以设计用于构建数据库索引的作为选择的方法。尤其是，下面所列许多步骤是可选的，但是用于产生更有效搜索的数据库索引。虽然搜索效率对于从大型数据库进行实时声音识别

很重要，但是，小型数据库可以相对快地搜索，即使其没有被最优地分类。

为索引声音数据库，集合中的每个记录都经受标志处理和指纹处理分析，来为每个音频文件产生一个索引集合。图4概略地说明了已经计算了标志(LM)和指纹(FP)的声音记录的片断。标志在声音的特定的时间点出现，并具有从文件的开头偏移的时间单位的值，而指纹在特定的标志处或其附近描述声音的特征。这样，在本实施例中，针对特定的文件的每个标志都是唯一的，而相同的指纹却可以在单个文件或多个文件内出现许多次。

在步骤42，使用在声音记录内发现与众不同且可再生的位置的方法，对每个音乐记录作标志。优选的标志处理算法能够在声音记录中标明相同的时间点，而不管噪声和其它线性及非线性失真的存在。某些标志处理方法在概念上独立于下述的指纹处理过程，但其可以被选择以优化其性能。标志处理导致声音记录中的一列时间点{$landmark_k$}，随后在这些时间点计算指纹。好的标志处理方案在声音记录中每秒标明大约5-10个标志；当然，标志密度依赖于声音记录中的活动(activity)的量。

多种技术可用于计算标志，其都在本发明的范围之内。用来实现本发明的标志处理方案的详细技术处理是本领域所公知的，故不再详细讨论。一种简单的标志处理技术被公知为功率范数(Power Norm)，在记录中的每一个可能的时间点处计算瞬时功率，并选择局部最大值。这样做的一种方式是通过对波形进行直接校正并滤波来计算包络。另一种方式为计算信号的希尔伯特变换(积分)，并使用希尔伯特变换和原始信号的平方值的和。

标志处理的功率范数方法长于发现声音信号中的瞬变。功率范数实际上是更一般的频谱Lp范数在p=2时的特殊情况。一般的频谱Lp范数是通过计算短时频谱，而沿声音信号的每一时刻计算的，例如，通过Hanning-windowed快速傅立叶变换(FFT)。一个优选实施例使用8000Hz的采样速率、1024个样本的快速傅立叶变换帧尺寸、以及每个时间段64个样本的步幅。然后计算频谱分量的绝对值的p次方之和作为针对每个时间段的Lp范数，可选地，再求p次方根。如前所述，在时间上选择结果值的局部最大值作为标志。图5示出了频谱Lp范数方法的一个例子，即针对特定的声音信号的L4范数作为时间的函数的图。局部最大值处的虚线表明所选标志的位置。

当p= 时，L 范数实际上是最大值范数。也就是说，范数的值是频谱段中最大频谱分量的绝对值。该范数带来强壮的(robust)标志和良好的整体识别

性能，并最好是用于音调的(tonal)音乐。

作为选择，通过在固定的或彼此可变的偏移处，求多个时间段上的频谱分量的绝对值 p 次方之和，来计算“多段”频谱标志，而不是单个段。发现该扩展的总和的局部最大值，允许多段指纹的位置的最优化，如下所述。

一旦计算了标志，在步骤 44 中，在记录中的每个标志时间点处计算指纹。一般地，指纹是概括在记录中该时间点处或其附近的一组特征的一个或一组值。在当前的优选实施例中，每个指纹是单个数字值，其为多个特征的隐(hashed)函数。指纹的可能类型包括频谱段指纹、多段指纹、线性预测编码系数、以及倒频谱系数。当然，任何类型的、描述信号或标志附近信号的特征的指纹都在本发明的范围之内。可以通过对信号的任何类型的数字信号处理或频率分析，来计算指纹。

为产生频谱段指纹，在每个标志时间点的附近执行频率分析，以抽取最高的几个频谱尖峰。简单的指纹值正好是最强频谱尖峰的单个频率值。使用这样的简单的尖峰，带来在存在噪声的情况中的令人惊讶的良好识别；然而，与其它指纹方案相比，单频谱段指纹往往产生更多的假正值(false positive)，因为其不是唯一的。可以通过使用由两个或三个最强频谱尖峰的一个函数构成的指纹，来减少假正值的数目。然而，如果第二强频谱尖峰不够强，不足以从存在的噪声中的竞争者中识别出，那么就可能对噪声更敏感。也就是说，所计算指纹值可能不够强壮，而不能可靠地再现。尽管如此，这种情况的性能也是好的。

为了利用许多声音的时间演化，通过向标志时间点添加一组时间偏移，来确定一组时间段。在每个所得时间段，计算频谱段指纹。然后组合所得的这组指纹信息，以形成一个多频声(multitone)或多段指纹。每个多段指纹远比单频谱段指纹更独特，因为，其跟踪时间演化，带来在下述的数据库索引搜索中的假匹配较少。实验表明，由于其增强的独特征，从两个时间段中的每一个中的单个最强频谱尖峰计算出的多段指纹，带来在随后的数据库索引搜索中快得多的计算(大约快 100 倍)，但是当存在显著的噪声时，识别百分率有一些下降。

作为选择，若不使用固定的便置或来自给定的时间段的偏移来计算多段指纹，则可以使用可变的偏移。对所选择的段的可变的偏移是，指纹从“锚(anchor)”标志到下一个标志、或到一定的偏移范围内的标志的偏移。在这种

情况中，标志之间的时间差值，连同多频率信息，也被编码到指纹中。通过向指纹添加更多维数，它们就会变得更加独特，且具有更低的假匹配的机会。

除频谱分量之外，可以抽取其它频谱特征，并用作指纹。线性预测编码分析，线性地抽取信号的诸如频谱尖峰、以及频谱形装之类的可预测特征。线性预测编码是数字信号处理领域中所公知的。对于本发明，通过将已量化的线性预测编码系数隐藏(hashing)进索引值中，可以将锚在标志位置处的波形段的线性预测编码系数用作指纹。

倒频谱系数在测量周期性时有用，并且可以被用于描述诸如语音或许多乐器之类的谐和的信号。倒频谱分析是数字信号处理领域中所公知的。对于本发明，许多倒频谱系数被一起隐藏进索引中，并用作指纹。

图 6 中示出了一个作为选择的实施例 50，在其中，同时计算标志和指纹。图 3 的步骤 42 和 44 被步骤 52、54、和 56 所取代。如下所述的，在步骤 52 中，从声音记录计算多维函数，并从该函数中抽取标志 54 和指纹 56。

在图 6 的实施例的一种实现中，从声音记录的声频谱图中计算标志和指纹。声频谱图是声音记录的时间-频率分析，在所述的声音记录中，对声音样本的窗口的(windowed)且重叠的帧做声频谱分析，典型地，使用快速傅立叶变换。如前所述，一个优选实施例使用 8000Hz 的采样速率、1024 个样本的快速傅立叶变换帧尺寸、以及每个时间段 64 个样本的步幅。图 7A 中示出了频谱图的一个例子。时间在水平轴上，而频率在垂直轴上。每个连续的快速傅立叶变换帧沿水平轴以相对应的等距间隔垂直堆叠。声频谱图描绘每一时间频率点的能量密度；图中较黑的区域代表较高的能量密度。声频谱图是音频信号处理领域中所公知的。对于本发明，可以从多个凸点中获取标志和指纹，所述的凸点如图 7B 的声频谱图中圈出的声频谱图局部最大值。例如，获取了每个尖峰的时间和频率坐标，其中，时间用作标志，而频率用来计算相对应的指纹。这种频谱图尖峰标志与 L 范数相似，在 L 范数中，由范数的最大绝对值确定标志位置。然而，在该声频谱图中，局部最大值搜索在时间-频率平面的斑点上进行，而不是在整个时间段上进行。

在本文中，将从声音记录的点抽取分析中而得来的凸点的集合称为星座(constellation)。对于由局部最大值构成的星座，优选分析为选择多个点，所述的多个点是每个所选点附近的时间-频率平面的最大能量值。例如，如果坐标(t_0, f_0)处的一个点在一个矩形内是最大能量值点，就选择坐标(t_0, f_0)处的点，

其中，所述的矩形的角坐标为(t_0-T, f_0-F)、(t_0-T, f_0+F)、(t_0+T, f_0-F)、以及(t_0+T, f_0+F)，即边长为 2T 和 2F 的矩形，而 T 和 F 被选择来提供适当数目的星座点。也可以根据频率值改变矩形的范围的大小。当然可以使用任何的区域形状。还可以对最大能量值标准加权，这样，竞争时间-频率能量尖峰根据时间-频率平面中的距离量度(metric)而被逆加权，即越远的点加权越小。例如，能量可以被加权为：

$$\frac{S(t,f)}{1+C_t(t-t_0)^2+C_f(f-f_0)^2},$$

其中，S(t, f)是声频谱图在点(t, f)处的幅度(magnitude)平方值，而 C_t 和 C_f 是正数值(不必是常数)。也可以是其它距离加权函数。局部最大值选择约束可以应用到其它(非最大值)凸点特征抽取方案，且在本发明的范围之内。

本方法带来与上述的单频谱指纹非常相似的、有着许多相同的属性的值对。声频谱图时间-频率方法比单频方法产生更多标志/指纹对，但是在下述的匹配阶段也可以得到许多假匹配。然而，其比单频谱指纹提供更强壮的标志处理和指纹处理，因为可以使声音样本中的强势噪声不扩展到每一段中声频谱的所有部分。也就是说，在声频谱的多个部分中，非常有可能有某些标志和指纹对没有被强势噪声所影响。

声频谱图标志处理和指纹处理方法是特征分析方法的特殊情况，所述的特征分析方法计算声音信号的多维函数，并在函数值中确定凸点的位置，其中，有一维是时间。凸点可以是局部最大值、局部最小值、零交叉(zero crossings)、或其它与众不同的特征。标志被作为凸点的时间坐标，而从其余的坐标中的至少一个来计算相对应的指纹。例如，多维凸点的非时间坐标可以隐藏(hashed)在一起，以形成多维函数的指纹。

上述的用于多段频谱指纹的可变的偏移方法可被应用于声频谱图或其它多维函数指纹。在这种情况中，如图 7C 中所示的声频谱图中所说明的，星座中的点被链接在一起而形成链接的点。星座中的每个点用作定义标志时间的锚点，其它点的其余坐标值被结合以形成链接的指纹。例如，彼此接近的点，如下所定义，被连接在一起形成更复杂的聚合体(aggregate)特征指纹，其可以更容易地被区分和搜索。和用多段频谱指纹一样，将信息从多链接的凸点结合到单个指纹中的目的是创建更多多样性的可能的指纹值，从而减少假匹配的可能性，即，减少用相同的指纹描述两个不同的音乐样本的可能性。

在原理上，在两点连接方案中，N 个凸点的每一个都可以链接到每一个其它点，产生大约 $N^2/2$ 个组合。相似地，对于 K 点连接，从一个星座引起的可能的组合的数目的量级是 N^K。为了避免这样的组合的激增，期望能约束将要连接在一起的点，使之相邻。完成这一约束的一种方式是为每个锚点定义一个“目标区域”。然后一个锚点与其目标区域中的多个点相连接。也可以选择目标区域内的点的子集来连接——并非每一个点都需要被连接。例如，只可以连接与目标区域中最强尖峰相关联的点。目标区域可以具有固定的形状，或根据锚点的特征而改变。对于声频谱图尖峰星座的锚点(t_0, f_0)的目标区域的简单例子是：使得 t 在间隔[t_0+L, t_0+L+W]中的声频谱图带中的点(t, f)的集合，其中，L 是进入将来的引子(lead)，而 W 是目标区域的宽度。在这种方案中，在目标区域中允许所有的频率。L 或 W 可以是变量，例如，如果使用一种比率控制机制调整所产生的连接组合的数目。作为选择，例如，通过约束目标区域使得为频率 f 在间隔[f_0-F, f_0+F]中，可以实现频率限制，其中，F 为边界参数。频率约束的一个优点在于：在心理声学中，已知当多个序列的音调具有彼此接近的频率时，旋律往往更好地一致。这样的约束可以使更多的“心理声学上逼真的”识别性能成为可能。尽管为心理声学建模不是本发明的必要目的。也可以考虑相反的规则，其中，f 被选为在区域[f_0-F, f_0+F]之外。这迫使连接彼此频率不同的点，可以避免下述情况，即星座抽取人为因素产生结结巴巴(stuttering)的、时间接近且频率相同的、多个序列的时间-频率点。如其它位置参数那样，F 不必是常数，并可以，例如，是 f_0 的函数。

当在指纹值中包括非锚凸点的时间坐标时，必须使用相对时间值，以允许指纹为时间不变量。例如，指纹可以是(i)非时间坐标值和/或(ii)多个凸点的相对应的时间坐标值的差值的函数。可以使用时间差值值，例如，关于锚点的，或在链接的集中的相继的凸点之间的连续差值。可以将坐标和差值放在链接的比特域(concatenated bit field)中，以形成隐藏的(hashed)指纹。由于本领域中的一般技术人员将会明白，存在将多组坐标值映射到指纹值的许多其它方式，且都在本发明的范围之内。

这种方案的一个具体例子使用 N>1 个链接的声频谱图尖峰，其坐标为(t_k, f_k)，k=1,…,N。然后，(i)取第一个尖峰的时间 t1 为标志时间，以及(ii)时间差值 $t_k=t_k$-t1,k=2,…,N，加上链接的尖峰的频率 f_k，k=1,…N，被隐藏(hashed)在一起以形成指纹值。指纹可以从所有可用的 t_k 和 f_k 坐标或其子集计算出。

例如，如果需要，可以忽略某些或所有时间差值坐标。

使用多点形成指纹的另一个优点在于，可以使指纹编码相对时间展宽不变，例如，当以不同于原始记录速度的速度播放声音记录时。这一优点既适用于声频谱图，又适用于时间段方法。注意到，在已展宽时间的信号中，时间差值值和频率具有反比关系(例如，以因数二减少时间差值，会使得频率加倍)。这种方法通过从指纹中移除时间展宽的方式将时间差值和频率进行结合，来利用了那种事实。

例如，在坐标值为(t_k, f_k)，k=1,…,N 的 N 点声频谱尖峰的情况中，将要隐藏(hash)进指纹中的可用的中介值(intermediate value)是 $t_k=t_k-t1,k=2,\ldots,N$，和 f_k，k=1,…N。然后，通过取多个频率中的一个，比如说 f_1，作为参考频率，并形成(i)其与其余频率的商、和(ii)其与时间差值的乘积，可以使中介值关于时间展宽不变。例如，中介值可以是 $g_k=f_k/f_1$，k=2,…,N，和 $s_k= t_kf_1$，k=2,…,N。如果样本以因数 α 加速，那么频率 f_k 变成 αf_k，而时间差值 t_k 变成 t_k/α，这样 $g_k=\alpha f_k/\alpha f_1=f_k/f_1$，而 $s_k=(t_k/\alpha)(\alpha f_1)= t_kf_1$。然后，使用函数将这些新中介值结合起来以形成独立于时间展宽的隐藏的(hashed)指纹值。例如，可以通过将 g_k 和 s_k 值放入链接的比特域中来隐藏(hash) g_k 和 s_k 值。

作为选择，可以使用参考时间差值，例如 t_2 来取代参考频率。新的中介值被计算为(i)与其余时间差值的商 t_k/ t_2 以及(ii)与频率的乘积 t_2f_k。这种情况等价于使用参考频率，因为结果值可以从上面的 g_k 和 s_k 值的乘积以及商求出。频率比率的倒数同样可以有效地被使用；也可以用原始中介值的对数值的和与差分别代替积与商。任何通过这样的换算(commutation)、代换(substitution)、以及置换(permutation)的数学操作所获取的时间展宽独立的指纹值都在本发明的范围之内。另外，可以使用多个参考频率或参考时间差值，它们也使时间差值相对化。使用多个参考频率或参考时间差值等价于使用单个参考值，因为可以通过对 g_k 和 s_k 值的算数操作实现相同的结果。

返回到图 3 和图 6，通过上述方法的任何一个进行标志处理和指纹处理分析会带来针对每个声音_ID 的索引集合，如图 8A 所示。针对给定的声音记录的索引集合是一列值对(指纹，标志)。典型地，每个被索引的记录在其索引集合中有一千的量级的(指纹，标志)对。在上述的第一个实施例中，标志处理和指纹处理技术基本上是独立的，可以视其为分离的且可交换的模块。按照系统、信号品质、或将要被识别的声音的类型，可以使用许多不同的标志处

理或指纹处理模块中的一个。事实上，因为索引集合简单地由多个值对组成，所以可以，而且往往最好是同时使用多个标志处理和指纹处理方案。例如，一种标志处理和指纹处理方案可能长于探测独特的音调模式，但是不长于辨认打击乐，因为不同的算法可能有相反的属性。使用多个标志处理/指纹处理策略带来更强壮且更丰富的识别性能的范围。通过为某些种类的指纹保留某些范围的指纹值，可以一起使用多种不同的指纹技术。例如，在32位指纹值中，可以用前3位限定后面的29位编码的是8个指纹处理方案中的哪一个。

为将要在声音数据库中索引的每个声音记录产生索引集合之后，以允许快速(即对数时间)搜索的方式构建可搜索的数据库索引。这是在步骤46中通过构建一列三元组(指纹，标志，声音_ID)来完成的，所述的三元组是通过向每个索引集合中的每个偶对(doublet)添加相对应的声音_ID而获取的。针对所有声音记录的所有这些三元组被收集到大型索引列表中，图8B中示出了其示例。然后，为了使随后的搜索处理最优化，根据指纹对该列三元组进行分类。快速分类算法是本领域中所公知的，而且，在D. E. Knuth, The Art of Computer Programming(计算机编程的技术)，Volume 3: Sorting and Searching(分类与搜索)，Reading，Massachusetts: Addison-Wesly，1998中被广泛地讨论过，在此并入作为参考。可以使用高性能分类算法在NlogN时间内对列表进行分类，其中，N是列表中的项目的数目。

一旦索引列表被分类，在步骤48中通过分段将其进行进一步处理，这样，列表中每个独特的指纹被收集到新的主索引列表，图8C中示出了它的一个例子。主索引列表中的每一个项目都包含指纹值和指向一列(标志，声音_ID)对的指针。按照被索引的记录的数目和特征，给定的指纹可以在整个集合中出现几百次甚至更多。将索引列表重新安排为主索引列表是可选的，但是节省存储器，因为每个指纹值只出现一次。其也可以加速随后的数据库搜索，因为列表中的项目的有效的数目极大地减少为一列独特的值。作为选择，可以通过将每个三元组插入一个B-tree(B树)来构建主索引列表。如本领域中的一般技术人员所公知的，存在用于构建主索引列表的其它可能性。主索引列表最好保留在诸如动态随机存取存储器(DRAM)之类的系统存储器中，用于在信号识别期间快速访问。主索引列表可以保留在系统内的单个节点的存储器中，如图2中所说明的。作为选择，主索引列表可以被分割成分配到多个计算节点中的块。参考上文的声音数据库索引最好是图8C中所说明的主索引列表。

声音数据库索引最好是离线(offline)构建，并且当识别系统中并入新的声音时，就增加地更新。为更新列表，可以向主列表中的适当的位置插入新的指纹。如果新的记录包含多个现有的指纹，那么向用于这些指纹的现有的列表添加相对应的(标志，声音_ID)对。

识别系统

使用如上所述地产生的主索引列表，对外来的声音样本执行声音识别，典型地，所述的声音样本是由希望辨认该样本的用户所提供的。例如，用户在广播上听到一首新歌曲，并想了解该歌曲的作者和名称。该样本可以源自诸如无线广播、迪斯科舞厅、酒馆、海底、声音文件、音频流片段、或立体声系统之类的任何类型的环境，并且可以包含背景噪声、信息遗失、或谈话语音。在向系统提供音频样本以供识别之前，用户可以将其存储在诸如应答机、计算机文件、磁带录音机、或电话或移动电话语音邮件系统之类的存储设备中。基于系统设置和用户约束，音频样本从诸如立体声系统、电视、光盘播放器、无线广播、应答机、电话、移动电话、因特网(Internet)流广播、文件传输协议(FTP)、作为电子邮件附件的计算机文件、或传送这样的记录材料的任何其它合适的装置之类的任意数目的模拟或数字来源提供给本发明的识别系统。按照来源，样本的形式可以是声波、无线电波、数字音频脉冲编码调制(PCM)流、压缩的数字音频流(诸如杜比数字(Dolby Digital)或运动画面专家组 3(MP3))、或因特网流广播。用户通过诸如电话、移动电话、网络浏览器、或电子邮件之类的标准接口与识别系统进行交互。样本可以被系统捕获并且实时处理，或者其可以被复制，用于从先前捕获的声音(例如声音文件)进行处理。在捕获期间，音频样本被数字地采样，并通过诸如麦克风之类的采样设备，将其发送到系统。按照捕获方法，样本可能会因信道或声音捕获设备的局限而经受进一步的劣化。

一旦声音信号被转换成数字形式，其被处理以便识别。如用于数据库文件的索引集合的构建，使用与用于处理声音记录数据库的算法相同的算法，来为样本计算标志和指纹。如果对原始声音文件的高度失真的再现的处理之后，能得到与针对原始记录所获得的相同的或相似的一组标志和指纹对，那么，该方法是最优的。针对声音样本的作为结果的索引集合是一组经分析的值对(指纹，标志)，如图 9A 中所示。

给定针对声音样本的多个对，搜索数据库索引以确定潜在匹配的文件的

位置。搜索按如下进行：通过在主索引列表中搜索 $fingerprint_k$，来处理未知的样本的索引集合中的每个($fingerprint_k$，$landmark_k$)对。关于有序的列表的快速搜索算法是本领域中所公知的，并且，在 D. E. Knuth，The Art of Computer Programming(计算机编程技术)，Volume 3: Sorting and Searching(分类与搜索)，Reading，Massachusetts: Addison-Wesly，1998 中被广泛地讨论过。如果在主索引列表中发现了 $fingerprint_k$，那么，其相对应的一列匹配的($landmark^*_j$，$sound\_ID_j$)对被复制，并增补 $landmark_k$，以形成形式为($landmark_k$，$landmark^*_j$，$sound\_ID_j$)的一组三元组。在这种符号表示法中，星号(*)表明数据库中的被索引的一个文件的标志，而没有星号的标志指的是样本。在某些情况中，最好是，匹配的指纹不需要是相同的，只需要是相似的；例如，在预先确定的阈值内，它们可以是不同的。匹配的指纹，不论是相同的还是相似的，都被称为是等价的。三元组中的 $sound\_ID_j$ 与具有带星号的标志的文件相对应。这样，每个三元组包含两个不同的标志，一个在数据库索引中，而一个在样本中，在这两个不同的标志处计算出等价的指纹。对所输入的样本的索引集合范围内的所有的 k 重复这种过程。将所有得到的三元组收集到一个大的候选列表中，如图 9B 中所说明的。称其为候选列表是因为：其包含多个声音文件的声音_ID，通过它们的匹配的指纹的特点，所述的声音文件是用于辨认外来的声音样本的候选者。

在编辑了候选列表之后，对其进一步的处理是根据声音_ID 分段。做这件事情的一种方便的方式是通过声音_ID 对候选列表进行分类，或将其插入到 B-树。如上所述，在本领域中有大量的分类算法可用。该处理的结果是一列候选声音_ID，其中，每一列都具有由样本和文件标志时间点对($landmark_k$，$landmark^*_j$)组成一个散布列表，其中，可选地剥去了声音_ID，如图 9C 中所示。这样，每一个散布列表包含一组相对应的标志，是根据它们的以等价的指纹值来描述的特征而相对应的。

然后分析针对每个候选声音_ID 的散布列表，以确定该声音_ID 是否与样本匹配。可以使用一个可选的阈值(thresholding)步骤，首先排除具有非常小的散布列表的潜在的大量的候选者。很明显，在其散布列表中只有一个项目的候选者，即只有一个指纹与样本一样的候选者，不与样本匹配。可以使用任何大于或等于一的合适的阈值数目。

一旦确定了候选者的最终的数目，就确定了选中的候选者的位置。如果

下面的算法不能确定选中的候选者的位置，则返回失败消息。洞察匹配处理的关键在于：假设两边的时间基(timebase)都是稳定的，则在匹配的声音中的时间演化必须遵循线性一致。这几乎总是正确的，除非一个声音已经被故意地非线性地扭曲了，或经受了诸如具有颤抖(warbling) 速度问题的盒式录放机之类的有缺陷的播放设备。这样，给定的声音_ID 的散布列表中的正确标志对($landmark_n$，$landmark^*_n$)必须有如下形式的线性一致：

$$landmark^*_n = m*landmark_n + offset,$$

其中，m 是斜率，应该接近一；$landmark_n$ 是外来的样本内的时间点；$landmark^*_n$ 是通过声音_ID 索引的声音记录内的相对应的时间点；而 offset 是偏移到与外来的声音样本的开头相对应的声音记录中的时间。能满足针对 m 和 offset 的特定的值的上述方程的多个标志对被称为线性相关。显然，线性相关的概念只对多于一对的相对应的标志有效。注意到，这种线性相关性以很高的概率辨认正确的声音文件，同时排除没有重要性的无关标志对。尽管对于两个不同的信号可以包含许多一致的指纹，但是这些指纹非常不可能具有相同的相关的(relative)时间演化。对线性一致的要求是本发明的关键特征，并提供一种识别技术，其显著地优于简单地计数相同的特征的数目或测量特征之间的相似性之类的技术。事实上，由于本发明的这一方面，即使在外来的声音样本中出现的原始记录的指纹少于 1%，即，如果声音样本非常短，或如果其是显著地失真的，仍然可以识别声音。

这样，确定是否有针对外来的样本的匹配的问题，被简化成等价于在给定的散布列表的标志点的散布图内发现斜率接近一的对角线。图 10A 和图 10B 中示出了两个样本散布图，其中，声音文件标志在水平轴上，而外来的声音样本标志在垂直轴上．在图 10A 中，辨认了斜率近似等于一的一条对角线，表明该歌曲确实与该样本匹配，即，该声音文件是选中的文件。水平轴上的截距表明偏移到该音频文件中，样本在那里开始。在图 10B 的散布图中，没有发现统计上有意义的对角线，表明该声音文件与外来的样本不匹配。

有很多种在散布图中发现对角线的方法，所有这些方法都在本发明的范围之内。可以理解，短语“确定对角线的位置”指的是等价于确定对角线的位置而又不明显地产生对角线的所有方法。一种优选的方法开始于：从上述方程的两端减去 $m*landmark_n$，将得到：

$$(landmark^*_n - m*landmark_n) = offset。$$

假设 m 近似等于一，即，假设没有时间展宽，我们可以得到：

$(landmark^*_n - landmark_n) = offset$。

然后，对角线发现(diagonal-finding)问题，被简化为发现针对给定的声音_ID 的、集群(cluster)在相同的 offset 值附近的多个标志对。这一点可以通过从一个标志减去另一个并收集所得偏移值的柱状图而容易地完成。可以通过使用快速分类算法对所得的偏移值进行分类、或通过创建具有计数器的箱(bin)项目并插入到 B-树中，来准备该柱状图。柱状图中的选中的偏移箱包含最大数目的点。在这里，这个箱被称为柱状图的尖峰。因为，如果外来的声音信号完全包含在正确的库(library)声音文件之中，则偏移必须为正，所以，可以排除导致负偏移的标志对。类似地，也可以排除超出文件的结尾的偏移。为每个有资格的声音_ID 记录在柱状图的选中的偏移箱内的点的数目。这个数目成为针对每个声音记录的分值。选择候选列表中的具有最高分值的声音记录为选中者。如下所述地，向用户报告选中的声音_ID，以发出通知辨认成功的信号。为防止辨认失败，可以使用最小阈值分值以控制辨认处理的成功。如果没有分值超过阈值的库声音，那么，就没有识别，并如此通知用户。

如果外来的声音信号包含多个声音，则可以识别每个单独的声音。在这种情况中，在校准扫描中确定多个选中者的位置。不需要知道声音信号包含多个选中者，因为校准扫描将确定分值远高于其余分值的多于一个的声音_ID 的位置。所使用的指纹方法最好展示出良好的线性重合(superposition)，以便可以抽取多个单独的指纹。例如，声频谱图指纹处理方法展示出线性重合。

如果声音样本已经经受了时间展宽，则斜率不是一致地等于一。对以时间展宽的样本(假设指纹是时间展宽不变量)假设一致的斜率的结果是：所计算出的偏移值不是相等的。解决这一问题并提供适度的时间展宽的方式是增加偏移箱的尺寸，即，考虑为在一个偏移范围内，是相等的。通常，如果多个点不落在一条直线上，则所计算的偏移值是显著地不同的，而在偏移箱的尺寸上的轻微增加并不会产生显著数目的假正值。

其它线发现(line-finding)策略是可以的。例如，可以使用 T. Risse，“Hough Transform for Line Recognition(用于线识别的 Hough 变换)”，Computer Vision and Image Processing(计算机视觉和图像处理)，46，327-345，1989 中描述的 Radon 或 Hough 变换，它们是机器视觉和图形研究领域中所公知的。在 Hough 变换中，散布图中的每个点在(斜率，偏移)空间中投影到一条直线。这样，在

Hough 变换中，散布图中的该组点被投影到多条直线的对偶空间(dual space)中。Hough 变换中的尖峰与参数直线的交叉点相对应。给定的散布图的这样的一个变换的全体的尖峰，表明 Hough 变换中交叉直线的最大数目，以及协同线性(co-linear)点的最大数目。为允许 5%的速度变化，例如，可以将 Hough 变换的结构限制到斜率参数在 0.95 与 1.05 之间变化的区域，这样，节省一些计算量。

分级搜索

除了排除具有非常小的散布列表的候选者的阈值步骤，还可以进一步地提高效率。在一个这样的提高中，根据出现的概率，数据库索引被分段成至少两部分，并且，起初只搜索具有匹配样本的最高概率的声音文件。该划分可以出现在处理的各种阶段。例如，主索引列表(图 8C)可以分段为两个或更多个部分，这样，步骤 16 和 20 首先在一个段上执行。也就是说，只从数据库索引的一部分中检索与匹配的指纹相对应的文件，并从这一部分中产生一个散布列表。如果没有确定选中的声音文件的位置，则对数据库索引的其余部分重复该处理。在另一个实现中，从数据库索引中检索所有的文件，但是对角线扫描在不同的段上分离地执行。

使用这种技术，首先在数据库索引中的声音文件的小子集上执行对角线扫描，所述的对角线扫描为本方法的计算量密集的部分。因为对角线扫描具有关于被扫描的声音文件的数目近似成线性的时间分量，执行这样的分级搜索很有优势。例如，假设声音数据库索引包含代表 1,000,000 个声音文件的指纹，但是只有大约 1000 个文件以很高的频率匹配样本查询，例如 95%的查询是针对 1000 个文件的，而只有 5%的查询是针对其余的 999,000 个文件的。假设计算成本与文件的数目是线性相关的，则成本在 95%的时间与 1000 成比例，而只有 5%的时间与 999,000 成比例。从而平均成本大约与 50,900 成比例。这样，分级搜索使计算量减小为二十分之一。当然，也可以将数据库索引分割成多于两级，例如一组新发行的歌曲、一组最近发行的歌曲、和一组老的不流行的歌曲。

如上所述，首先对声音文件的第一子集，即高概率文件，执行搜索，然后，仅当首次搜索失败时，对包含其余文件的第二子集执行搜索。如果每个偏移箱中的点的数目没达到预定的阈值，则对角线扫描失败。作为选择，可以并行地(同时地)执行这两级搜索。如果在对第一子集的搜索中确定了正确的

声音文件的位置，则发出信号以终止对第二子集的搜索。如果没有在对第一个搜索中确定出正确的声音文件的位置，则继续第二个搜索，直到确定选中的文件的位置。这两种不同的实现涉及到在计算力(effort)和时间上的权衡。第一种实现有更好的计算效率，但是，如果第一个搜索失败了，就引入了轻微的等待时间；而对于第二种实现，如果选中的文件是在第一子集中，则会浪费计算力，但是，当选中的文件不是在第一子集中时，等待时间最小。

对列表分段的目的是估计一个声音文件为查询的目标的概率，并将搜索限制在那些最有可能与查询样本匹配的文件中。有各种可能的方式来指定概率并对数据库中的声音分类，它们全部都在本发明的范围之内。最好是，基于新旧程度和被识别为选中的声音文件的频率来指定概率。新旧程度是有用的量度，特别是对于流行歌曲，因为随着新歌曲的发行，音乐兴趣随时间变化非常快。在计算概率分值之后，为文件指定等级，而列表按等级自分类(self-sort)。然后，已分类的列表被分段成两个或更多个子集，用于搜索。较小的子集可以包含预定的数目的文件。例如，如果排列将文件定位于顶部，换句话说，1000 个文件中，则将文件放置在较小的、更快的搜索中。作为选择，可以动态调整用于两个子集的分界点。例如，所有具有超过一个特定的阈值的分值的文件可以被放置在第一子集中，并因此，每个子集中的文件的数目连续变化。

计算概率的一种特定的方式是：每当一个声音文件被辨认为针对查询样本的一个匹配时，就给声音文件的分值增加一。为说明新旧程度，周期性地降低所有记录的分值，这样，较新的查询比较旧的查询在排列上具有更强的效果。例如，可以通过对每个查询使用常数乘法因数，来变(ratchet)低所有的分值，使得：如果未被更新，分值就呈指数衰减。按照数据库中的文件的数目，该数目很容易达到一百万，这种方法要求每次查询时更新大量的分值，使得其潜在地不受欢迎。作为选择，以相对不频繁的间隔向下调整分值，比如每天一次。从较不频繁的调整得到的顺序，与从每次查询时都调整得到的顺序，是有效地相似的，但不十分一致。然而，更新等级的计算量非常低。

这种新旧程度调整的一个轻微变化是：每当查询时，向选中的声音文件添加指数增长的分值更新 a^t，其中，t 是从上次全体更新以来所经过的时间的量，该变化更准确地保持新旧程度分值。然后在每次全体更新时，通过将所有的分值除以 a^T，来向下调整所有的分值，其中，T 是从上次全体更新以来

所经过的总的时间。在这中变化中，a是大于一的新旧因数。

除了上述的排列，可以引入某些推理知识，以帮助选择列表中的种子候选者(seed)。例如，新发行的歌曲很可能比旧歌曲拥有更高的查询数。这样，新发行的歌曲可以被自动地放置在第一子集中，所述的第一子集包含具有匹配查询的较高概率的歌曲。这可以独立于上述的自排列算法而被执行。如果也使用自排列特征，新发行的歌曲可以被指定初始等级，以将其放置在第一子集中的某个地方。新发行的歌曲可以被种子候选(seed)在列表的非常顶部的位置、高概率歌曲列表的底部、或两者之间的某个地方。由于搜索的目的，初始位置并不重要，因为等级将随时间而收敛，以反映真实的兴趣水平。

在一个作为选择实施例中，搜索是以新旧排列的顺序来执行的，并在声音_ID值超过预定的阈值时终止。这与每段只包含一个声音_ID的上述的方法是等价的。实验表明，选中的声音的分值远大于所有其它声音文件的分值，并因此可以用最小限度的实验来选择合适的阈值。实现本实施例的一种方式是：根据新旧程度排列数据库索引中的所有的声音_ID，并对相同分值的情况中进行任意的重新评比(tie-breaking)。因为每个新旧程度排列是唯一的，所以新旧程度分值与声音_ID之间是一一映射。于是，当按声音_ID分类时，可以使用排列来代替声音_ID，以形成候选声音_ID 的列表和关联的散布列表(图9C)。在产生三元组(指纹，标志，声音_ID)的索引列表时，索引列表被分类为主索引列表之前，可以将排列号限定(bound)在索引中。然后，以排列代替声音_ID。作为选择，可以使用搜索和替换函数来用排列替换声音_ID。只要保持映射完整性，随着排列被更新，新的排列就被映射到旧的排列上。

作为选择，在处理中，排列可以在稍后被限定(bound)。一旦创建了散布列表，排列可以与每个声音_ID相关联。然后，通过排列对多个集进行分类。在这个实现中，只需要修改指向散布列表的指针；不需要重复分组成散布列表。稍后限定(bindings)的优点是：不需要在每次更新排列时重建整个数据库索引。

注意到，流行度(popularity)排列本身就可以作为有经济价值的对象。也就是说，排列映用户获取一个未知的声音样本的辨认的需求。在很多情况中，查询是由购买该歌曲的欲望引起的。实际上，如果已知关于用户的人口统计信息，那么可以为每个期望的人口统计群体实现作为选择的排列方案。可以从用户签订识别服务时要求的简介信息获取用户的人口统计群体。也可以通

过标准协作滤波技术(standard collaborative filtering technique)动态地确定。

在实时系统中，声音随时间递增地提供给识别系统，使得能够流水(pipelined)识别。在这种情况中，可以分段处理输入数据，并递增地更新样本索引集合。在每次更新周期之后，使用上述的搜索和扫描步骤，最近增加的索引集合被用来检索候选库声音记录。从数据库索引中搜索与最近获取的样本指纹匹配的指纹，并产生新的($landmark_k$，$landmark^*_j$，$sound\_ID_j$)三元组。散布列表中添加了新的对，且柱状图也被增加。这一途径的优点在于：如果已经收集了可以毫不含糊地辨认声音记录的充足数据，例如，如果多个声音文件中的一个的偏移箱中的点的数目超过一个高阈值或超过第二高的声音文件分值，那么就可以终止数据采集并宣布结果。

一旦辨认了正确的声音，用任何合适的方法向用户或系统报告结果。例如，结果可以通过计算机打印输出、电子邮件、网络搜索结果页面、发给移动电话的短消息服务 (SMS)、通过电话的计算机产生的语音通知、或将结果公布到用户可以稍后访问的网站或因特网帐号上。所报告的结果可以包括诸如歌曲的名称和作者、古典作品的作曲家和名称以及记录属性(例如，演奏者、指挥、演出地)、广告的公司和产品、或任何其它合适的标识符之类的声音的辨认信息。另外，可以提供传记信息、关于附近音乐会的信息、和其它歌迷感兴趣的信息；可以提供到这类数据的超级链接。所报告的结果也可以包括声音文件的绝对分值、或其与第二高分值的文件相比的分值。

本识别方法的一个有用的成果在于：其不混淆相同声音的两个不同的表演。例如，古典音乐的同一篇章的不同演奏不会被认为是相同的，即使人们不能察觉到两者之间的差异。这是因为，针对两次不同的演奏的标志/指纹对与其时间演化，极不可能精确地匹配。在当前的实施例中，标志/指纹对彼此必须在 10 毫秒之内，以便辨认为线性一致。作为结果，本发明所执行的自动识别使得在所有情况下都可以信任合适的演奏/声迹(soundtrack)和作者/标签。

实现的例子

下面描述本发明的一个优选的实现，即连续滑动的窗口音频识别。麦克风或其它声源被连续地采样到缓冲器中，以获取声音的前 N 秒记录。周期性地分析声音缓冲器的内容，以确定声音内容的一致性。声音缓冲器可以具有固定的尺寸，或可以随着声音被采样而增长尺寸，在这里，被称为音频样本的顺序增长段。给出报告以表明被辨认的声音记录的出现。例如，可以收集

日志文件，或在设备上显示指示诸如标题、艺术家、唱片封面画、歌词、或购买信息之类的关于音乐的信息。为避免冗余，可以只在所识别的声音的一致性改变时给出报告，例如，在自动唱片点唱机节目改变之后。这样的设备可以被用于创建从任何声音流(无线广播、因特网信息流广播、隐藏的麦克风、电话呼叫、等)播放的音乐的列表。除了音乐一致性，可以把诸如识别的时间之类的信息记入日志。如果位置信息是可获得的(例如，从全球定位系统(GPS))，也可以把这类信息记入日志。

为完成该辨认，可以重新辨认每个缓冲器。作为选择，例如，可以将声音参数抽取到指纹或其它中间的特征抽取的形式中，并存储在第二个缓冲器中。新指纹可以被添加到第二个缓冲器的前端，并从该缓冲器的尾端丢弃旧指纹。这样的一个滚动缓冲器方案的优点在于，不需要对声音样本的重叠的旧段冗余地执行相同的分析，这样，来节省计算力。对滚动指纹缓冲器的内容周期性地执行辨认处理。在小型便携式设备的情况中，由于指纹流不是非常数据密集的，所以，可以在设备中执行指纹分析，且可以使用相对低带宽的数据信道将结果传送给识别服务器。滚动指纹缓冲器可以被保留在便携式设备中，并每次向识别服务器传输，或可以保留在识别服务器上，这种情况中，连续识别会话(session)被高速缓存(cache)在服务器上。

在这样的一个滚动缓冲器识别系统中，一有充分的信息可用于识别，就可以识别新的声音记录。充足的信息可以占用小于缓冲器的长度。例如，如果一首与众不同的歌曲在播放一秒钟后就可以唯一的识别，并且系统识别周期为一秒钟，那么，就可以立即识别该歌曲，尽管缓冲可以有 15-30 秒长。相反地，如果一首较少特色的歌曲要求更多秒钟的样本来识别，那么，在宣布歌曲的一致性之前，系统就必须等待较长的时期。在这种滑动窗口识别方案中，一有声音可以被辨认，就可以识别该声音。

需要非常注意的是，尽管已经以完整功能的识别系统和方法描述了本发明，那些本领域中的技术人员将会明白，本发明的机制能够以各种形式的指令的计算机可读媒体的形式而被分配，并且，本发明可以平等地应用，而不管用于实际执行分配的信号承载媒体的具体类型。这类计算机可存取设备的例子包括计算机存储器(随机存取存储器或只读存储器 (ROM))、软盘、和光盘只读存储器(CD-ROM)，以及诸如数字和模拟通讯链接之类的传输类型的媒体。

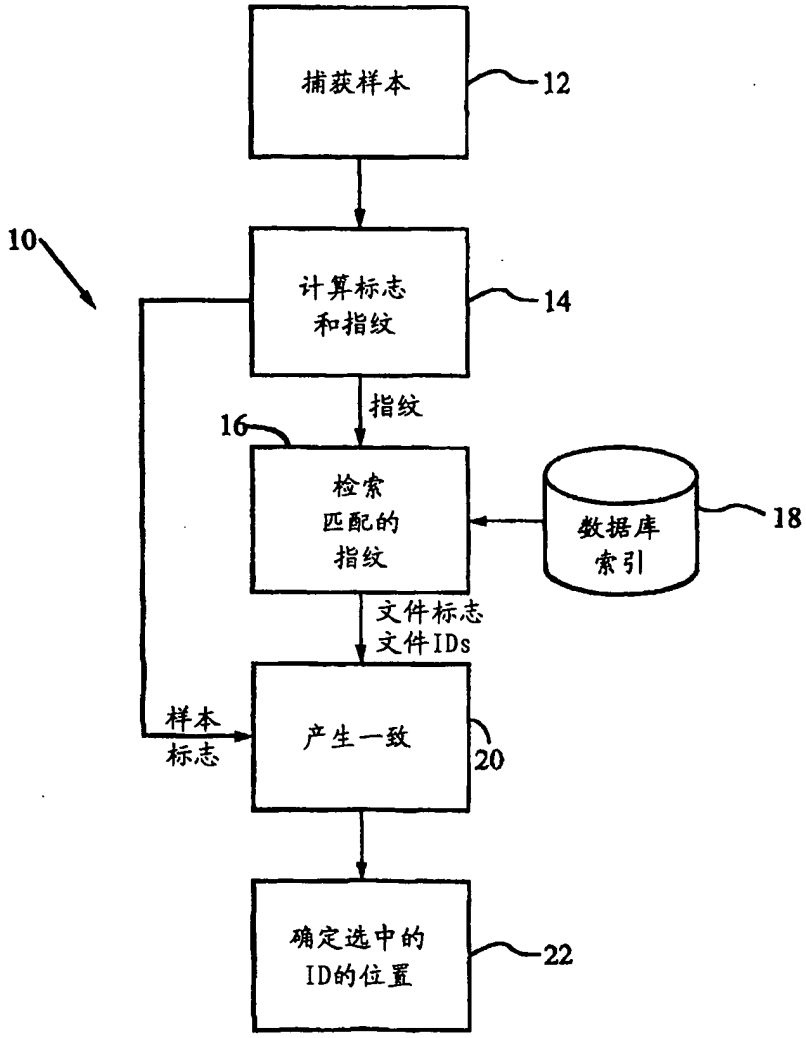

捕获样本
12
10
计算标志
和指纹
14
指纹
16
检索
匹配的
指纹
数据库
索引
18
文件标志
文件IDs
样本
标志
产生一致
20
确定选中的
ID的位置
22

图 1

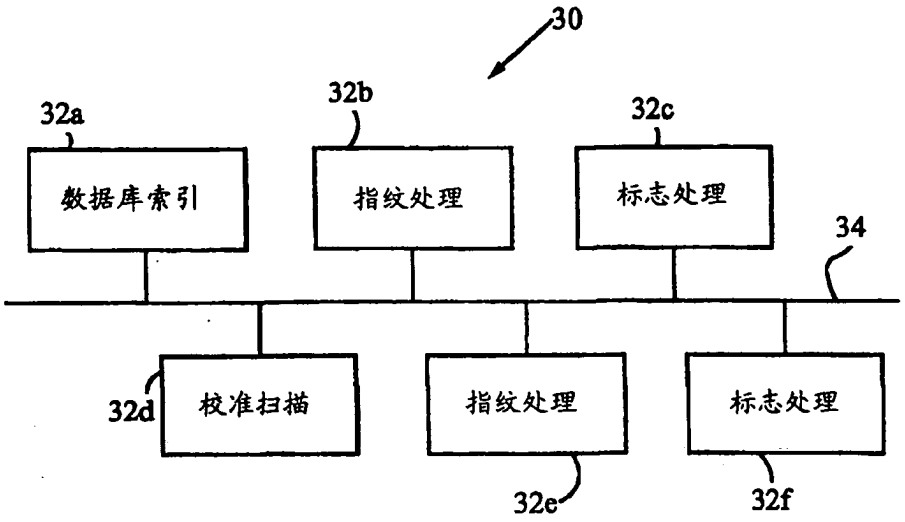

30
32a
数据库索引
32b
指纹处理
32c
标志处理
34
32d
校准扫描
指纹处理
标志处理
32e
32f

图 2

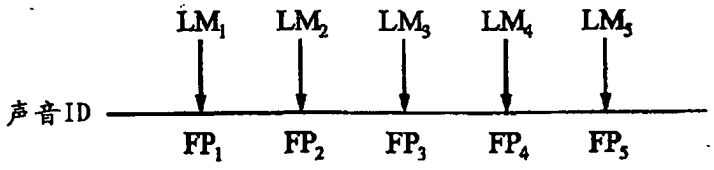

LM1
LM2
LM3
LM4
LM5
声音ID
FP1
FP2
FP3
FP4
FP5

图 4

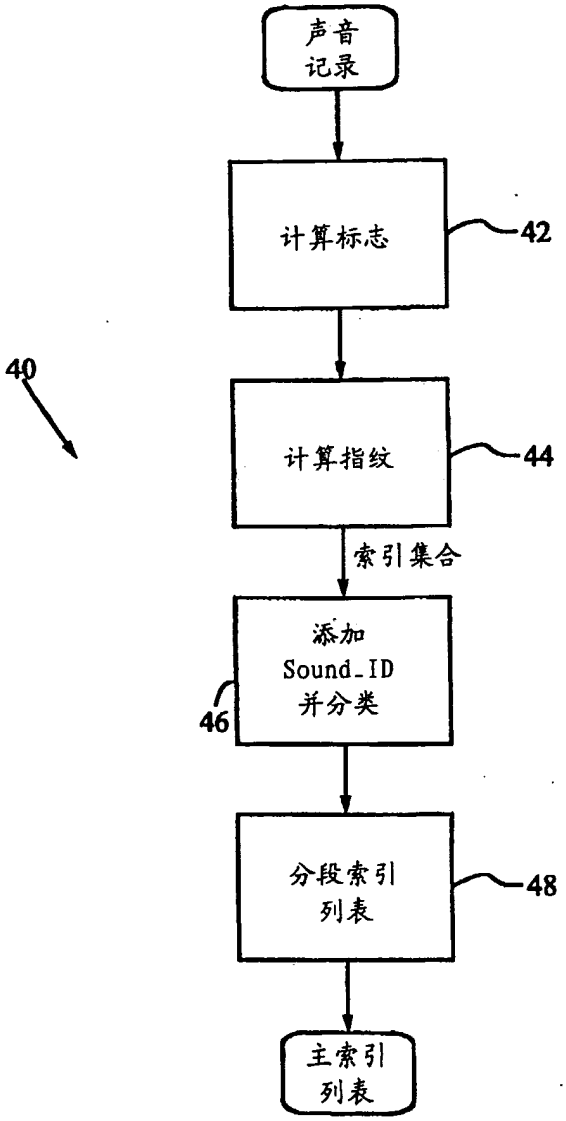


图 3

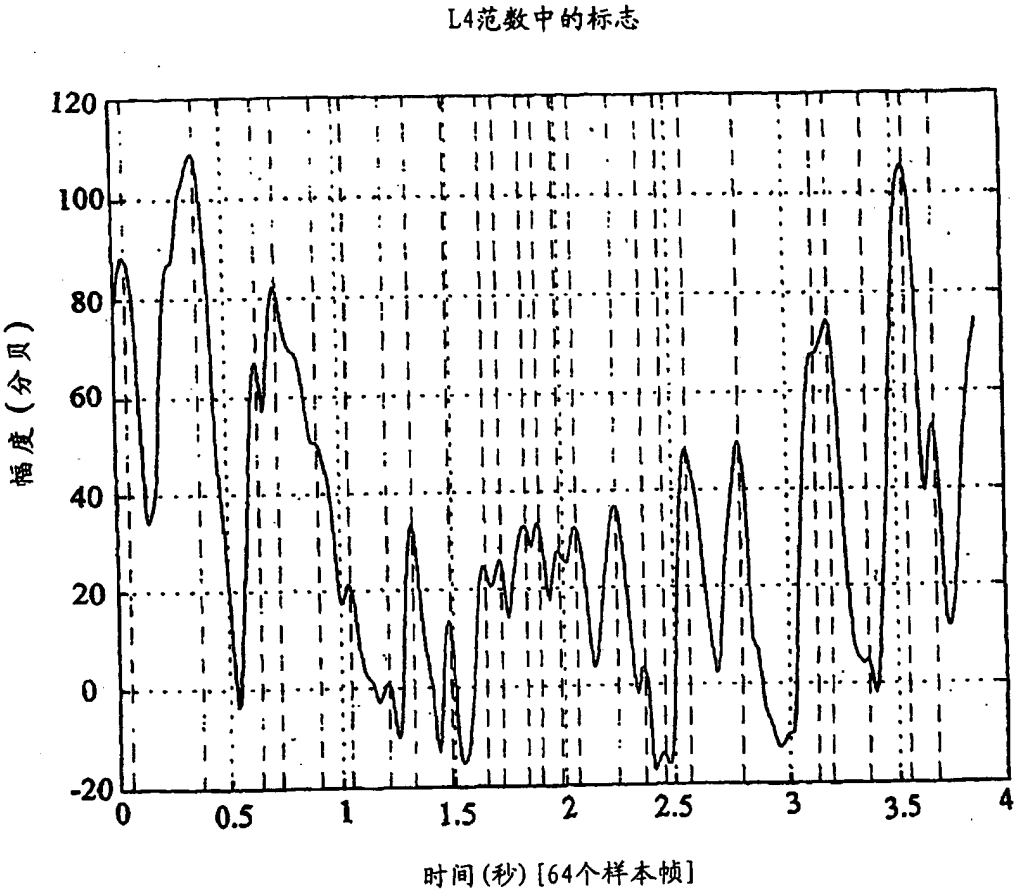

L4范数中的标志
120
100
80
60
40
20
0
-20
0
0.5
1
1.5
2
2.5
3
3.5
4
幅度(分贝)
时间(秒)[64个样本帧]

图 5

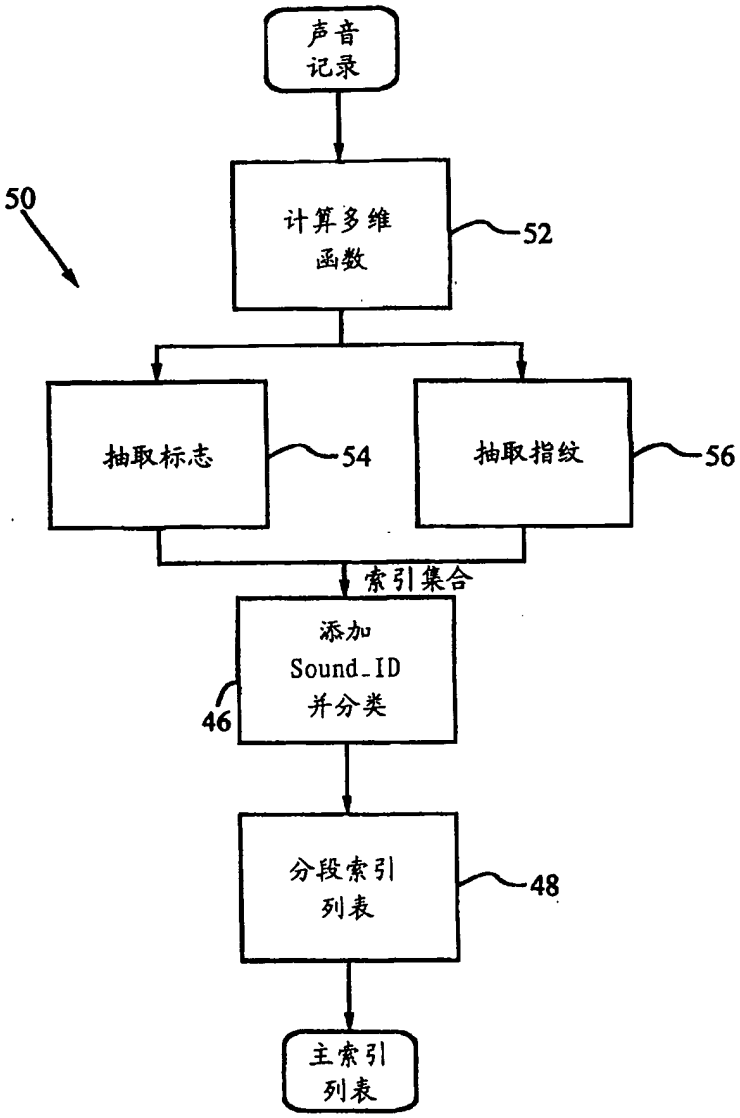

声音
记录
50
计算多维
函数
52
抽取标志
54
抽取指纹
56
索引集合
添加
Sound_ID
并分类
46
分段索引
列表
48
主索引
列表

图 6

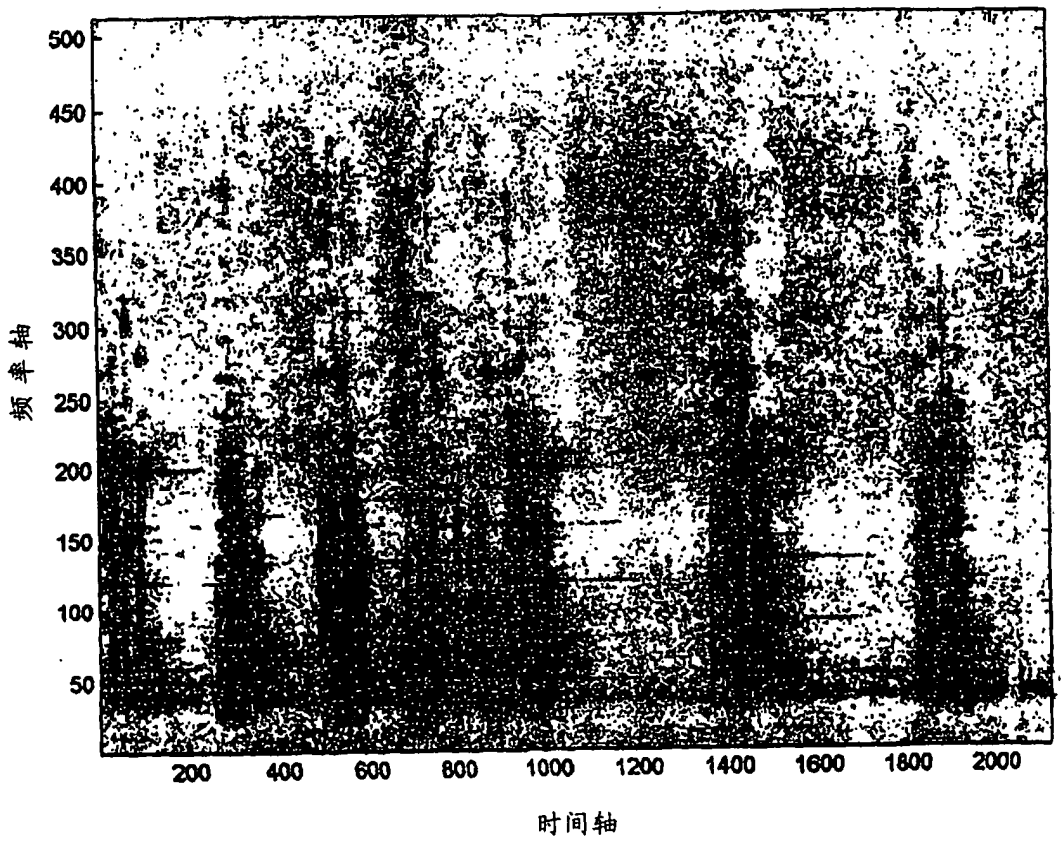

500
450
400
350
300
250
200
150
100
50
频率轴
200
400
600
800
1000
1200
1400
1600
1800
2000
时间轴

图 7A

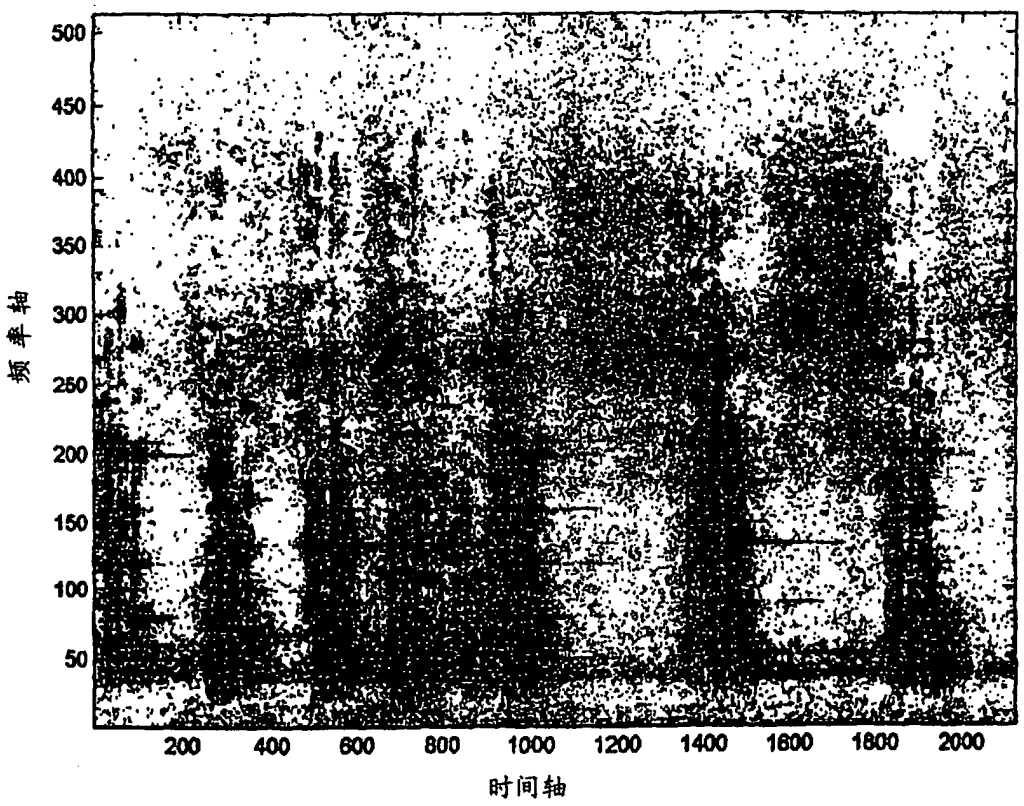

500
450
400
350
300
250
200
150
100
50
频率轴
200
400
600
800
1000
1200
1400
1600
1800
2000
时间轴

图 7B

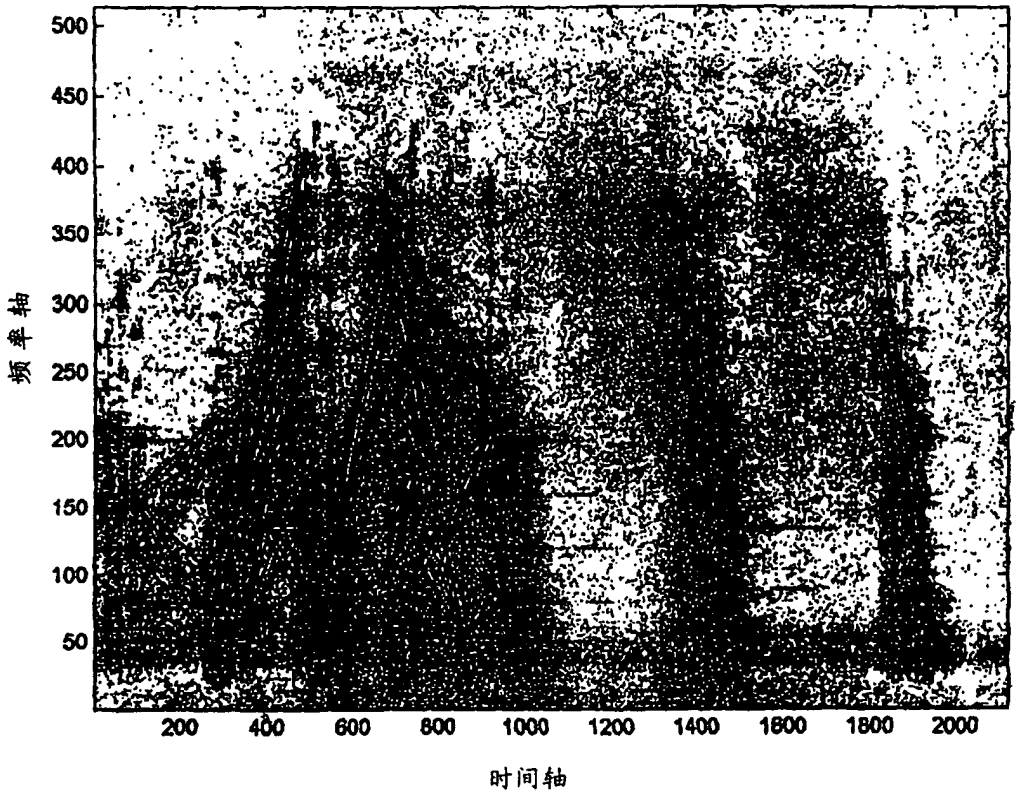

500
450
400
350
300
250
200
150
100
50
频率轴
200
400
600
800
1000
1200
1400
1600
1800
2000
时间轴

图 7C

索引集合

$$\begin{matrix} \text{Sound\_ID}_1 & & \text{Sound\_ID}_n \\ \begin{bmatrix} (\text{fingerprint}_1, \text{landmark}_1) \\ (\text{fingerprint}_2, \text{landmark}_2) \\ \vdots \\ (\text{fingerprint}_m, \text{landmark}_m) \end{bmatrix} & \cdots & \begin{bmatrix} (\text{fingerprint}_p, \text{landmark}_p) \\ (\text{fingerprint}_{p+1}, \text{landmark}_{p+1}) \\ \vdots \\ (\text{fingerprint}_t, \text{landmark}_t) \end{bmatrix} \end{matrix}$$

图 8A

索引列表

$$\begin{bmatrix} (\text{fingerprint}_1, \text{landmark}_1, \text{sound\_ID}_1) \\ (\text{fingerprint}_2, \text{landmark}_2, \text{sound\_ID}_1) \\ \vdots \\ (\text{fingerprint}_m, \text{landmark}_m, \text{sound\_ID}_1) \\ \vdots \\ (\text{fingerprint}_p, \text{landmark}_p, \text{sound\_ID}_n) \\ (\text{fingerprint}_{p+1}, \text{landmark}_{p+1}, \text{sound\_ID}_n) \\ \vdots \\ (\text{fingerprint}_t, \text{landmark}_t, \text{sound\_ID}_n) \end{bmatrix}$$

图 8B

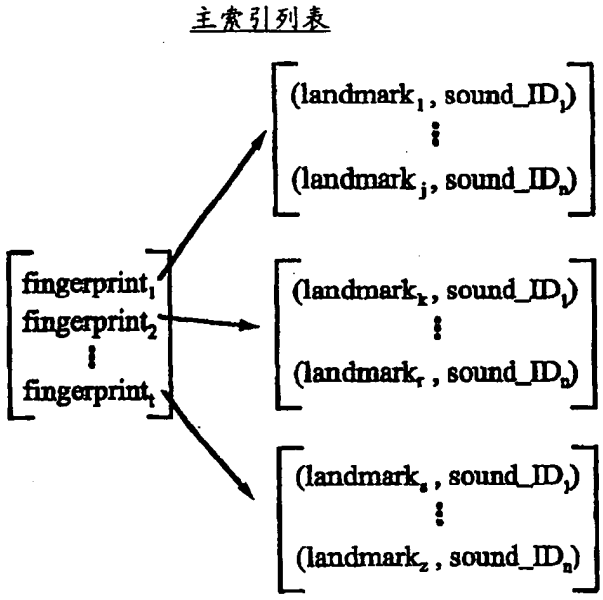

主索引列表
(landmark1, sound_ID1)
(landmarkj, sound_IDn)
fingerprint1
fingerprint2
fingerprintt
(landmarkk, sound_ID1)
(landmarkr, sound_IDn)
(landmarks, sound_ID1)
(landmarkz, sound_IDn)

图 8C

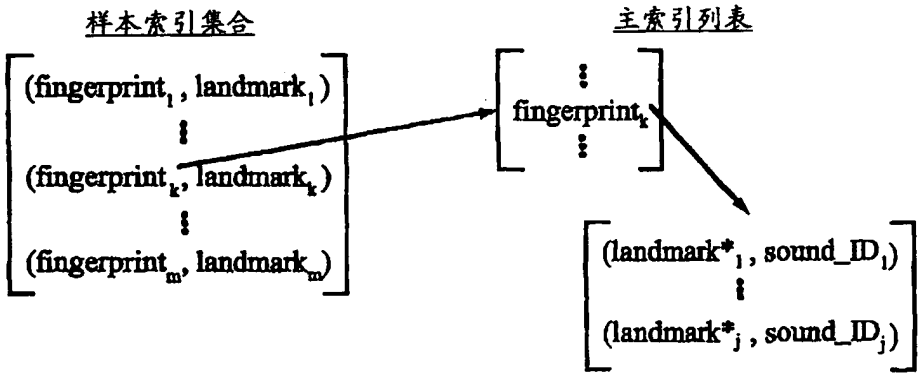

样本索引集合
(fingerprint1, landmark1)
(fingerprintk, landmarkk)
(fingerprintm, landmarkm)
主索引列表
fingerprintk
(landmark*1, sound_ID1)
(landmark*j, sound_IDj)

图 9A

候选列表

$$\begin{bmatrix} (landmark_1, landmark^*_1, sound\_ID_1) \\ \vdots \\ (landmark_k, landmark^*_j, sound\_ID_j) \\ \vdots \\ (landmark_m, landmark^*_r, sound\_ID_n) \end{bmatrix}$$

图 9B

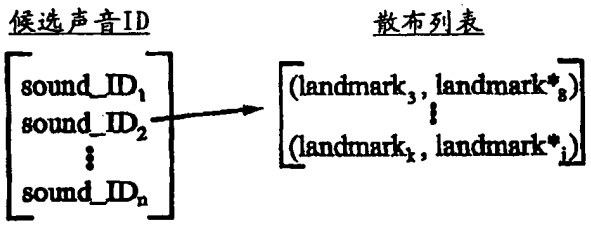


图 9C

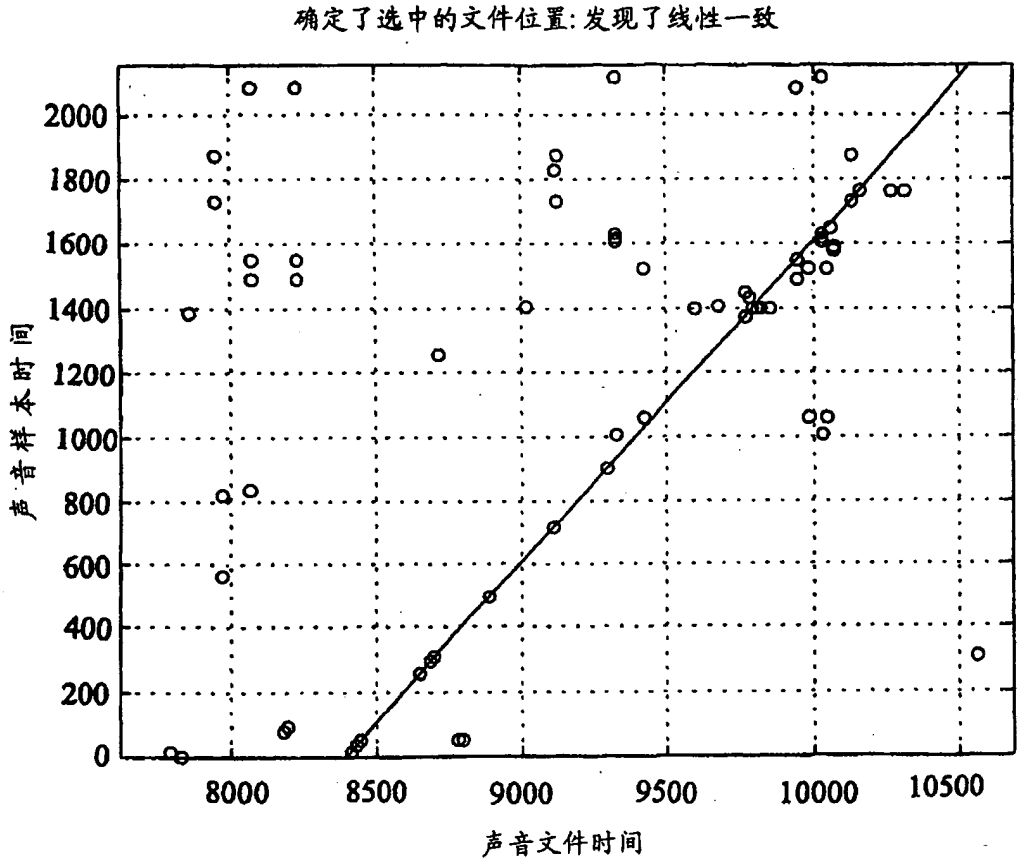

确定了选中的文件位置:发现了线性一致
2000
1800
1600
1400
1200
1000
800
600
400
200
0
声音样本时间
8000
8500
9000
9500
10000
10500
声音文件时间

图 10A

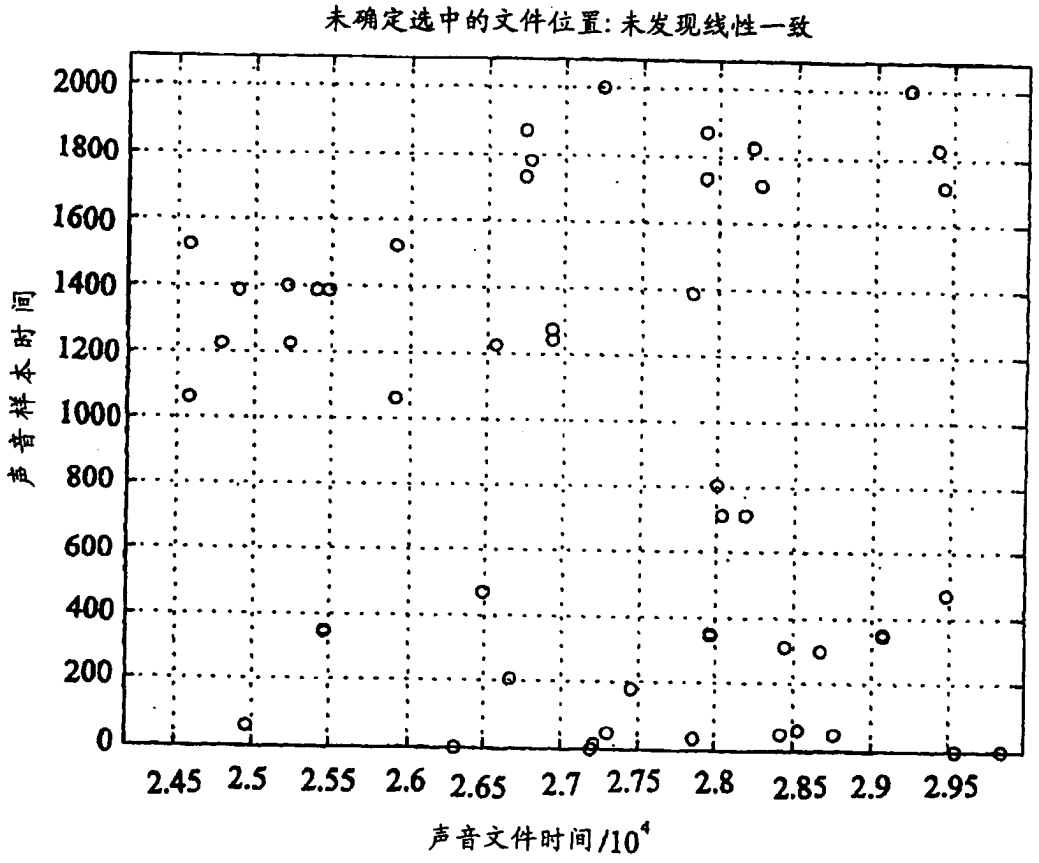

未确定选中的文件位置: 未发现线性一致
2000
1800
1600
1400
1200
1000
800
600
400
200
0
声音样本时间
2.45
2.5
2.55
2.6
2.65
2.7
2.75
2.8
2.85
2.9
2.95
声音文件时间/10^4

图 10B